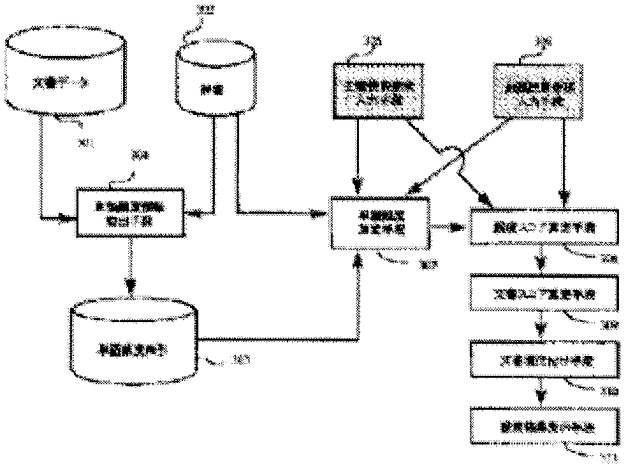
DOCUMENT RETRIEVING DEVICE

Publication number: JP10049549
Publication date: 1998-02-20
Inventor: INABA MITSUAKI; NOGUCHI NAOHIKO; SUGANO YUJI; SATO MITSUHIRO; NOMOTO MASAKO; YASUKAWA HIDEKI
Applicant: MATSUSHITA ELECTRIC IND CO LTD
Classification:
- international: ***G06F17/30;*** **G06F17/30;** (IPC1-7): G06F17/30
- European: G06F17/30T2P4V
Application number: JP19970087328 19970324
Priority number(s): JP19970087328 19970324; JP19960156418 19960529

Also published as:
EP0810535 (A2)
US6154737 (A1)
EP0810535 (A3)
CN1172994 (A)
EP0810535 (B1)
more >>

Abstract of **JP10049549**

PROBLEM TO BE SOLVED: To provide a document retrieving device capable of shortening the overall retrieving time including highly accurate retrieval and convergence. SOLUTION: A frequency score calculating means 308 calculates a frequency score indicating a matching degree between a document and a retrieval request by word frequency from the total number of documents, the number of documents in which a certain word appears, the appearance frequency of the word in each document, and the weighting parameter of the word which are outputted from a word frequency calculating means 307 and a document score calculating means 309 calculates a document score indicaing a matching degree between the document and the retrieving request from the frequency score and orders the score, so that a retrieving result more close to a user's retrieving intension can be obtained.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（A）　(11)特許出願公開番号

特開平10－49549

(43)公開日　平成10年(1998) 2月20日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/30			Ｇ０６Ｆ　15/403	３４０Ｂ	
			15/40	３７０Ａ	

審査請求　未請求　請求項の数７　ＦＤ　（全 27 頁）

(21)出願番号　特願平9－87328

(22)出願日　平成９年(1997) ３月24日

(31)優先権主張番号　特願平8－156418

(32)優先日　平８(1996) ５月29日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地

(72)発明者　稲葉　光昭
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　野口　直彦
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　菅野　祐司
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(74)代理人　弁理士　役　昌明　（外３名）

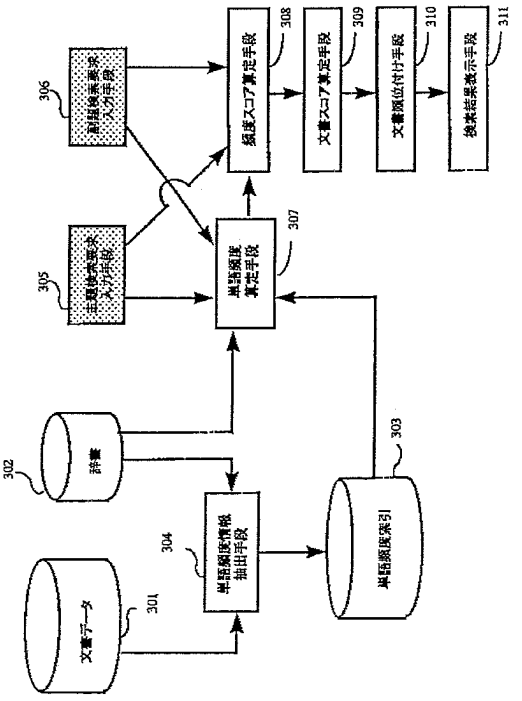
最終頁に続く

(54)【発明の名称】　文書検索装置

(57)【要約】

【課題】　文書データからユーザの入力した検索要求に合致する文書を探索し、その合致する度合によって順位付けを行なう文書検索装置に関するもので、従来の単語頻度のみによる文書の順位付けのもっていたユーザの検索意図に反した記事が上位に来てしまう問題点を解決し、高精度な検索と絞り込みを含めた総合的な検索時間の短縮を可能にする文書検索装置の提供を目的とする。

【解決手段】　頻度スコア算定手段308は単語頻度算定手段307が出力した総文書数、単語の出現する文書数、文書における単語の出現頻度、単語の重み付けパラメータから、単語頻度による文書と検索要求の合致度合を示す頻度スコアを算出し、文書スコア算定手段309が上記頻度スコアから文書と検索要求の合致度合を示す文書スコア算出し、順位付けを行なうことによって、ユーザの検索意図により近い検索結果を得ることが可能となる。



1

【特許請求の範囲】

【請求項１】　検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、複数の検索要求入力手段を備え、優先度の異なる複数の検索要求をユーザが入力できるようにしたことを特徴とする文書検索装置。

【請求項２】　検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書の複数のフィールドに対して、各々索引情報を持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段を備え、検索対象文書の順位付けに反映させる割合をフィールド毎にユーザが指定できるようにしたことを特徴とする文書検索装置。

【請求項３】　検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語数算定手段を備え、検索要求に含まれる複数の単語が検索対象文書中に同時に現れる場合に、当該文書に与える得点を加算することによりこれを優先的に表示させるようにしたことを特徴とする文書検索装置。

【請求項４】　検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語出現位置を索引に持ち、検索要求に含まれる複数の単語の検索対象文書中での出現位置の隣接度合を調べる単語近接度算定手段を備え、出現位置の近接度合によって、当該文書に与える得点を加算することによりこれを優先的に表示させるようにしたことを特徴とする文書検索装置。

【請求項５】　検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報を索引に持ち、複数の検索要求入力手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べる単語共起関係照合手段を備え、優先度の異なる複数の検索要求をユーザが入力できるようにすると共に単語共起関係が現れる文書に与える得点を加算することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置。

【請求項６】　検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報をフィールド毎に索引に持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べるフィールド別単語共起関係照合手段を備え、検索対象文書の順位付けに反映させる割合をフィールド毎にユーザが指定できるようにすると共にフィールド毎に単語共起関係が現れる文書に

2

与える得点を加算することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置。

【請求項７】　検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報を索引に持ち、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語算定手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べる単語共起関係照合手段を備え、検索要求に含まれる複数の単語が検索対象文書中に同時に現れる場合に、当該文書に与える得点を加算すると共に単語共起関係が現れる文書に与える得点を加算することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は文書検索装置に関し、特に大量の文書データからユーザの入力した検索要求に合致する文書を探索し、その合致する度合によって順位付けを行なう文書検索装置に関するものである。

【０００２】

【従来の技術】近年、文書検索の分野においては文書データベースの大規模化が進み、従来のようなキーワード検索や全文検索ではたとえ検索結果が高速に得られたとしても、その数が膨大で目的の文書を捜し出すのが困難な場合が増え、トータルな検索時間は必ずしも短縮されているとは言い難い。結果文書数を減らすためには、さらに別のキーワードを追加するなどして絞り込みを行なうという方法もあるが、目的とする文書が不必要な文書とともにふるい落とされてしまわないような適切なキーワードを追加するのは難しい。

【０００３】そこで、検索対象文書に文字列が存在するか否かだけでなく、その出現頻度等に着目して検索結果文書の順位付けを行ない、目的の文書を効率良く捜し出せるランキングの手法が注目されている。

【０００４】図１１は検索結果の順位付けを行なう従来の文書検索装置の構成を示したブロック図である。図１１において、文書検索装置は、検索対象となる文書データ1101と、辞書1102と、辞書単語の文書中での出現頻度を格納した単語頻度索引1103と、文書データ1101から単語の出現頻度情報を得る単語頻度情報抽出手段1104と、ユーザからの検索要求を入力する検索要求入力手段1105と、単語頻度索引1103から単語の出現頻度を求める単語頻度算定手段1106と、単語の出現頻度をもとに各文書の頻度スコアを求める頻度スコア算定手段1107と、頻度スコアをもとに各文書と検索要求の合致度合を示す文書スコアを算出する文書スコア算定手段1108と、文書スコアの順に文書を並べ替える文書順位付け手段1109と、スコアの順に並べられた結果文書を表示する検索結果表示手段1110とから構成されている。

【０００５】図１２は検索結果の順位付けを行なう従来の文書検索装置の索引作成の手順を示した流れ図である。まず、検索の前に予め単語頻度情報抽出手段1104が文書データ1101を調べ、総文書数、出現文書数とともに単語頻度索引1103に出力し単語頻度索引を作成しておく。

【０００６】ステップ１２０１において、検索するユーザは検索要求入力手段1105により、検索要求を入力する。ステップ１２０２において、単語頻度算定手段1106＊は単語頻度索引1103を参照し、総文書数ＮＤおよび、検索要求入力手段1105で入力された検索要求に含まれる辞書単語Ｗｉ（ｉ＝１、２、・・・、ＮＷ：ＮＷは検索要求に含まれる辞書単語数）について、当該単語の出現文書数ＮＤｉ、文書Ｄｊ（ｊ＝１、２、・・・、ＮＤ）中での出現頻度ＴＦｉｊを算出する。ステップ１２０３において、文書スコア算定手段1107が単語頻度算定手段1106からの出力をもとに文書Ｄｊの頻度スコアＳＦｊを式（１）によって算出する。

$$SFj=\sum_{i}(TFij\times IDFi)$$

$$IDFi=1-\log(NDi/ND) \quad \cdots(1)$$

ここで、ＩＤＦｉは単語Ｗｉの全文書における偏りを表すパラメータである。

【０００７】ステップ１２０４において、文書スコア算定手段1108が頻度スコア算定手段1107の出力した文書Ｄ※ｊの頻度スコアＳＦｊをもとに文書Ｄｊと検索要求の合致度合を示す文書スコアＳｊを求める。従来の検索装置においては式（２）のように文書スコアＳｊは頻度スコアＳＦｊそのものである。

$$Sj=SFj \quad \cdots(2)$$

【０００８】ステップ１２０５において、文書順位付け手段1109が文書スコア算定手段1108で算出された各文書のスコアの大きい順に検索結果を並べ替え、ステップ１２０６において、検索結果表示手段1110がその検索結果をユーザに提示する。

【０００９】

【発明が解決しようとする課題】しかしながら前記の従来の構成では、図１３に示すように、検索要求のうちの１つの単語が非常に高頻度で出現するような文書があった場合、ユーザの検索意図に反した文書であっても、高い順位にランクされてしまうという課題を有していた。また、検索対象文書の順位付けに用いるスコアの算出は、フィールドに関係なく文書単位で行なわれるため、その文書の内容を良く表しているであろうと思われる新聞記事の「見出し」や特許の「発明の名称」等の情報が活用できないという課題を有していた。

【００１０】また、複数の検索要求を与える場合、それらに優先順位をつけられず、ユーザの要求を柔軟に表現できないという課題や、全てを含んでいて欲しい単語群を検索要求として与えた場合でも、１つの単語が非常に高頻度で出現する文書があると高順位にきてしまうという課題や、近接して出現しなければ意味の無いような単語群を検索要求として表現し、検索することができないといった課題を有していた。

【００１１】本発明は前記従来技術の課題を解決するために、ユーザの検索要求を柔軟に受け付け、検索、順位付けを行なうことにより、ゴミの少ない高精度な検索結果が得られ、結果の絞り込みを含めた総合的な検索時間が短縮可能な文書検索装置を提供することを目的とする。

【００１２】

【課題を解決するための手段】本発明の文書検索装置においては、検索対象文書中の単語共起関係の情報を索引に持ち、検索要求に含まれる単語共起関係が検索対象文書に現れるか否かを調べる共起関係算定手段を備えたものである。

【００１３】また、優先度の異なる複数の検索要求を入力するために、複数の検索要求入力手段を備えたものである。

【００１４】また、検索対象文書の複数のフィールドに対して、各々索引情報を持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段を備えたものである。

【００１５】また、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語数算定手段を備えたものである。

【００１６】また、検索対象文書中の単語出現位置を索引に持ち、検索要求に含まれる複数の単語の検索対象文書中での出現位置が隣接もしくは比較的近いかどうかを調べる単語近接度算定手段を備えたものである。

【００１７】本発明によれば、ユーザの検索要求を柔軟に表現し、検索、順位付けを行なうことにより、ユーザの検索意図に沿った、ゴミの少ない高精度な検索結果が得られ、結果の絞り込みを含めた総合的な検索時間が短縮可能な文書検索装置が得られる。

【００１８】

【発明の実施の形態】本発明の請求項１記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、複数の検索要求入力手段を備え、優先度の異なる複数の検索要求をユーザが入力できるようにしたことを特徴とするものであり、ユーザの目的とする文書を高精度で検索することが可能となる作用を有する。

【００１９】また、本発明の請求項２記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書の複数のフィ

ールドに対して、各々索引情報を持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段を備え、検索対象文書の順位付けに反映させる割合をフィールド毎にユーザが指定できるようにしたことを特徴とするものであり、ユーザの目的とする文書を高精度で検索することが可能となる作用を有する。

【００２０】また、本発明の請求項３記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語算定手段を備え、検索要求に含まれる複数の単語が検索対象文書中に同時に現れる場合に、当該文書に与える得点を加算することによりこれを優先的に表示させるようにしたことを特徴とするものであり、ユーザの目的とする文書を高精度で検索することが可能となる作用を有する。

【００２１】また、本発明の請求項４記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語出現位置を索引に持ち、検索要求に含まれる複数の単語の検索対象文書中での出現位置の近接度合を調べる単語近接度算定手段を備え、出現位置の近接度合によって、当該文書に与える得点を加算することによりこれを優先的に表示させるようにしたことを特徴とするものであり、ユーザの目的とする文書を高精度で検索することが可能となる作用を有する。

【００２２】また、本発明の請求項５記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報を索引に持ち、複数の検索要求入力手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べる単語共起関係照合手段を備え、優先度の異なる複数の検索要求をユーザが入力できるようにすると共に単語共起関係が現れる文書に与える得点を加算することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置としたものであり、ユーザの目的とする文書をより高精度で検索することが可能となる作用を有する。

【００２３】また、本発明の請求項６記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報をフィールド毎に索引に持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べるフィールド別単語共起関係照合手段を備え、検索対象文書の順位付けに反映させる割合をフィールド毎にユーザが指定できるようにすると共にフィールド毎に単語共起関係が現れる文書に与える得点を加算することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置としたものであり、ユーザの目的とする文書をより高精度で検索することが可能となる作用を有する。

【００２４】また、本発明の請求項７記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報を索引に持ち、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語算定手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べる単語共起関係照合手段を備え、検索要求に含まれる複数の単語が検索対象文書中に同時に現れる場合に、当該文書に与える得点を加算すると共に単語共起関係が現れる文書に与える得点を加算することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置としたものであり、ユーザの目的とする文書をより高精度で検索することが可能となる作用を有する。

【００２５】以下、本発明の実施の形態について、図を参照しながら説明する。

【００２６】（第１の実施の形態）図１は本発明の第１の実施の形態における文書検索装置の構成を示したブロック図である。図１において、文書検索装置は、検索対象となる文書データ301と、辞書302と、辞書単語の検索対象文書中における出現頻度を格納した単語頻度索引303と、文書データ301から単語頻度情報を抽出し、単語頻度索引303に格納する単語頻度情報抽出手段304と、ユーザが優先させたいと思う検索要求を入力するための主題検索要求入力手段305と、主題検索要求に比べ優先度の低い検索要求をユーザが入力するための副題検索要求入力手段306と、主題検索要求入力手段305および副題検索要求入力手段306で入力された検索要求に含まれる辞書単語について、単語頻度索引303を調べ各文書中での出現頻度を求める単語頻度算定手段307と、単語頻度算定手段307で得られた単語頻度をもとに各文書の頻度スコアを求める頻度スコア算定手段308と、頻度スコア算定手段308の出力をもとに各文書の文書スコアを算出する文書スコア算定手段309と、スコアの順に文書を並べ替える文書順位付け手段310と、スコアの順に並べられた結果文書を表示する検索結果表示手段311とから構成されている。

【００２７】図２は本発明の第１の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【００２８】まず、検索の前に予め、単語頻度情報抽出手段304が文書データ301を走査し、辞書302に含まれる辞書単語の各文書中での出現頻度を調べ、総文書数、当該単語の出現文書数とともに単語頻度索引303に出力し、索引を作成しておく。

【００２９】ステップ４０１において、ユーザは主題検索要求入力手段305によって探したい文書に対する検索

要求でかつ重視してほしいものを主題検索要求として入力する。ステップ４０２において、ユーザは副題検索要求入力手段306によってそれほど重視しなくてもよいものを副題検索要求として入力する。

【００３０】ステップ４０３において、単語頻度算定手段307は単語頻度索引303を参照し、総文書数ＮＤを求めるとともに、主題検索要求入力手段305および副題検索要求入力手段306で入力された検索要求に含まれる辞書単語Ｗｉ（ｉ＝１、２、・・・、ＮＷ：ＮＷは検索要求に含まれる辞書単語数）に対し、当該単語の文書Ｄｊ（ｊ＝１、２、・・・、ＮＤ）中での出現頻度ＴＦｉｊ＊と当該単語の出現する文書数ＮＤｉを取得し、頻度スコア算定手段308に出力する。ステップ４０４において、単語頻度算定手段307は単語Ｗｉが主題検索要求、副題検索要求のどちらに含まれるかによってパラメータＷＴｉを選び、ステップ４０５において、頻度スコア算定手段308は単語頻度算定手段307が出力した総文書数ＮＤ、単語Ｗｉの出現する文書数ＮＤｉ、文書Ｄｊにおける単語Ｗｉの出現頻度ＴＦｉｊ、単語Ｗｉの重み付けパラメータＷＴｉから、単語頻度による文書Ｄｊと検索要求の合致度合を示す頻度スコアＳＦｊを式（３）によって算出する。

$$SFj=\sum_{i}(TFij\times IDFi\times WTi)$$

$$IDFi=1-\log(NDi/ND) \quad \cdots (3)$$

ここで、ＩＤＦｉは単語Ｗｉの全文書における偏りを表すパラメータである。

【００３１】ステップ４０６において、文書スコア算定手段309は頻度スコア算定手段308の出力した頻度スコアから文書Ｄｊと検索要求の合致度合を示す文書スコアＳｊを前記式（２）によって算出する。

【００３２】ステップ４０７において、文書順位付け手段310は文書スコア算定手段309の出力した各文書Ｄｊの文書スコアＳｊの値の大きい順に文書を並べ替え、ステップ４０７において、検索結果表示手段311が文書順位付け手段310の出力から得られたソート済の文書を検索の結果としてユーザに表示する。

【００３３】前記のようにして、ユーザが入力する検索要求に優先度を設けることにより、ユーザの検索意図を柔軟に表現することができ、効率的な検索が可能となる。

【００３４】（第２の実施の形態）図３は本発明の第２の実施の形態における文書検索装置の構成を示したブロック図である。図３において、文書検索装置は、検索対象となる文書データ501と、辞書502と、辞書単語の検索対象文書中における出現頻度を各フィールド毎に格納したフィールド別単語頻度索引503と、文書データ501から単語頻度情報を抽出し、フィールド別単語頻度索引503に格納する単語頻度情報抽出手段504と、ユーザが検索要求を入力するための検索要求入力手段505と、文書のどのフィールドのスコアをどの程度文書スコアに反映させるかという割合をユーザが入力するためのフィールド割合入力手段506と、検索要求入力手段505で入力された検索要求に含まれる辞書単語について、フィールド別単語頻度索引503を調べ、各文書中でのフィールド毎の出現頻度を求めるフィールド別単語頻度算定手段507と、フィールド別単語頻度算定手段507で得られた単語頻度をもとに各文書のフィールド別の頻度スコアを求めるフィールド別頻度スコア算定手段508と、フィールド別頻度スコア算定手段508の出力とフィールド割合入力手段506からの入力をもとに各文書の文書スコアを算出する文書スコア算定手段509と、スコアの順に文書を並べ替える文書順位付け手段510と、スコアの順に並べられた結果文書を表示する検索結果表示手段511とから構成されている。

【００３５】図４は本発明の第２の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【００３６】まず、検索の前に予め、単語頻度情報抽出手段504が文書データ501を走査し、辞書502に含まれる辞書単語の各文書内でのフィールド毎の出現頻度を調べ、総文書数、当該単語の出現文書数とともに出現頻度をフィールド別単語頻度索引503に出力し、索引を作成しておく。

【００３７】ステップ６０１において、ユーザは検索要求入力手段505によって探したい文書に対する検索要求を入力する。ステップ６０２において、ユーザはフィールド割合入力手段506によってフィールドＦｋ（ｋ＝１、２、・・・、ＮＦ：ＮＦは総フィールド数）のスコアを順位付けに反映させる割合Ｒｋを入力する。

【００３８】ステップ６０３において、フィールド別単語頻度算定手段507はフィールド別単語頻度索引503を参照し、総文書数ＮＤを求めるとともに、検索要求入力手段505によって入力された検索要求に含まれる辞書単語Ｗｉ（ｉ＝１、２、・・・、ＮＷ：ＮＷは検索要求に含まれる辞書単語数）に対し、当該単語の文書Ｄｊ（ｊ＝１、２、・・・、ＮＤ）のフィールドＦｋ中での出現頻度ＴＦｉｊｋとフィールドＦｋに当該単語の出現する文書数ＮＤｉｋを取得し、フィールド別頻度スコア算定手段508に出力する。

【００３９】ステップ６０４において、フィールド別頻度スコア算定手段508はフィールド別単語頻度算定手段507が出力した総文書数ＮＤ、フィールドＦｋに単語Ｗｉの出現する文書数ＮＤｉｋ、文書Ｄｊのフィールドにおける単語Ｗｉの出現頻度ＴＦｉｊｋ、単語頻度による文書Ｄｊのフィールドと検索要求の合致度合を示す頻度スコア頻度スコアＳＦｊｋを式（４）によって算出する。

9　　10

$$SFjk=\Sigma_{i}(TFijk\times IDFik)$$

$$IDFik=1-\log(NDik/ND) \quad \cdots (4)$$

【００４０】ステップ６０５において、文書スコア算定手段509はフィールド別頻度スコア算定手段508の出力したフィールド毎の頻度とスコアフィールド割合入力手段＊506で入力されたフィールドＦｋを反映させる割合Ｒｋから、文書Ｄｊと検索要求の合致度合を示す文書スコアＳｊを式（５）によって算出する。

$$Sj=\Sigma_{k}(SFjk\times Rk) \quad \cdots (5)$$

【００４１】ステップ６０６において、文書順位付け手段510は文書スコア算定手段509の出力した各文書Ｄｊの文書スコアＳｊの値の大きい順に文書を並べ替え、ステップ６０７において、検索結果表示手段511が文書順位付け手段510の出力から得られたソート済の文書を検索の結果としてユーザに表示する。

【００４２】前記のようにして、ユーザが検索対象フィールドのスコア配分の割合を変化させられるようにすることにより、ユーザの検索意図を柔軟に表現することができ、効率的な検索が可能となる。

【００４３】（第３の実施の形態）図５は本発明の第３の実施の形態における文書検索装置の構成を示したブロック図である。図５において、文書検索装置は、検索対象となる文書データ701と、辞書702と、辞書単語の検索対象文書中での出現頻度を格納した単語頻度索引703と、文書データ701から単語頻度情報を抽出し、単語頻度索引703に格納する単語頻度情報抽出手段705と、検索要求をユーザが入力するための検索要求入力手段707と、検索要求入力手段707で入力された検索要求に含まれる辞書単語について、単語頻度索引703を調べ当該単語の文書中での出現頻度を求める単語頻度算定手段708と、単語頻度算定手段708で得られた単語頻度をもとに各文書のスコアを求める頻度スコア算定手段709と、単語頻度索引703を調べ、検索要求入力手段707で入力された検索要求に含まれる単語のうちいくつが、文書中に出現するかを求める出現語数算定手段710と、出現語数算定手段710で得られた出現語数に基づいて各文書に加算するスコアを求める出現語数スコア算定手段711と、頻度スコア算定手段709および出現語数スコア算定手段711の出力から各文書のスコアを算出する文書スコア算定手段712と、スコアの順に文書を並べ替える文書順位付け手段713と、スコアの順に並べられた結果文書を表示する検索結果表示手段714とから構成されている。

【００４４】図６は本発明の第３の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【００４５】まず、検索の前に予め、単語頻度情報抽出手段705が文書データ701を走査し、辞書702に含まれる辞書単語の各文書内での出現頻度を調べ、総文書数、当該単語の出現文書数とともに出現頻度を単語頻度索引70※3に出力し、索引を作成しておく。

【００４６】ステップ８０１において、ユーザは検索要求入力手段707によって探したい文書に対する検索要求を入力する。検索要求は複数の単語を入力してもよいし、文章を入力し別途単語抽出手段を用いて文章から単語を切り出すようにしてもよい。

【００４７】ステップ８０２において、単語頻度算定手段708は単語頻度索引703を参照し、総文書数Ｎを求めるとともに検索要求入力手段707で入力された複数の辞書単語Ｗｉ（ｉ＝１、２、・・・、ＮＷ：ＮＷは検索要求に含まれる辞書単語数）に対し、当該単語の文書Ｄｊ（ｊ＝１、２、・・・、ＮＤ）中での出現頻度ＴＦｉｊと単語Ｗｉの出現する文書数ＮＤｉを取得し、頻度スコア算定手段709に出力する。

【００４８】ステップ８０３において、頻度スコア算定手段709は単語頻度算定手段708が出力した総文書数ＮＤ、単語Ｗｉの出現する文書数ＮＤｉ、文書Ｄｊにおける単語Ｗｉの出現頻度ＴＦｉｊから、単語頻度による文書Ｄｊと検索要求の合致度合を示す頻度スコアＳＦｊを前記式（１）によって算出する。

【００４９】ステップ８０４において、出現語数算定手段710はステップ８０２までで既に得られている文書Ｄｊに出現する辞書単語の情報と検索要求入力手段707で入力された検索要求に含まれる複数の単語Ｗｉを比較し、複数の単語Ｗｉのうちで文書Ｄｊに出現するものの数ＮＡｊを算出し、出現語数スコア算定手段711に出力する。

【００５０】ステップ８０５において、出現語数スコア算定手段711は出現語数算定手段710が出力した検索要求に含まれる単語のうちで文書Ｄｊに出現するものの数ＮＡｊに基づいた出現語数スコアＳＡｊを算出する。例えば式（６）によって算出することができる。

$$SAj=NAj-1 \quad \cdots (6)$$

【００５１】ステップ８０６において、文書スコア算定手段712は頻度スコア算定手段709が出力した頻度スコアＳＦｊと出現語数スコア算定手段711が出力した出現語数スコアＳＡｊから検索要求と文書Ｄｊの合致度合を表すスコアＳｊを式（７）によって算出する。

$$Sj=SFj+SAj\times 定数 \quad \cdots (7)$$

【００５２】出現語数スコアＳＡｊを用意することにより、検索要求に含まれる単語をより多く含むような文書のスコアを高くし、優先的に表示させることが可能となる。また、式（７）において定数の値を変化させること

により、出現語数による優先表示の度合を変えることも可能である。

【００５３】ステップ８０７において、文書順位付け手段713は文書スコア算定手段712が出力した各文書Ｄｊの文書スコアＳｊの値の大きい順に文書を並べ替える。ステップ ８０８において、検索結果表示手段714は文書順位付け手段713の出力から得られたソート済の文書を検索の結果としてユーザに提示する。

【００５４】前記のようにすれば、検索要求に複数の単語を含む場合に高頻度単語を１つだけ含むような文書が検索結果の上位に来てしまうというような不都合を回避でき、効率的な検索が可能となる。

【００５５】（第４の実施の形態）図７は本発明の第４の実施の形態における文書検索装置の構成を示したブロック図である。図７において、文書検索装置は、検索対象となる文書データ901と、辞書902と、辞書単語の検索対象文書中での出現頻度を格納した単語頻度索引903と、検索対象文書中に現れる単語の位置を格納した単語出現位置索引904と、文書データ901から単語頻度情報を抽出し、単語頻度索引903に格納する単語頻度情報抽出手段905と、文書データ901から単語の位置情報を求め、単語出現位置索引904に格納する単語出現位置情報抽出手段906と、検索要求をユーザが入力するための検索要求入力手段907と、検索要求入力手段907で入力された検索要求に含まれる辞書単語について、単語頻度索引903を調べ当該単語の文書中での出現頻度を求める単語頻度算定手段908と、単語頻度算定手段908で得られた単語頻度をもとに各文書のスコアを求める頻度スコア算定手段909と、単語出現位置索引904を参照し、検索要求入力手段907で入力された検索要求に含まれる単語の文書中での出現位置を求める出現位置算定手段910と、単語出現位置算定手段910の出力から単語どうしの近接度合を求める単語近接度算定手段911と、単語近接度算定手段911の出力に基づいて各文書に加算するスコアを求める近接スコア算定手段912と、頻度スコア算定手段909および近接スコア算定手段912の出力から各文書のスコアを算出する文書スコア算定手段913と、スコアの順に文書を並べ替える文書順位付け手段914と、スコアの順に並べられた結果文書を表示する検索結果表示手段915とから構成されている。

【００５６】図８は、本発明の第４の実施の形態における文書検索装置の検索の手順を示した流れ図である。ま＊ず、検索の前に予め、単語頻度情報抽出手段905が文書データ901を走査し、辞書902に含まれる辞書単語の各文書内での出現頻度を調べ、総文書数、当該単語の出現文書数とともに出現頻度を単語頻度索引903に出力し、単語出現位置情報抽出手段906が辞書単語の各文書中での出現位置を調べ、単語出現位置索引904に出力し、索引を作成しておく。

【００５７】ステップ１００１において、ユーザは検索要求入力手段907によって探したい文書に対する検索要求として複数の単語を入力する。なお、検索要求としてユーザは文章を入力し、別途単語抽出手段を用いて文章から単語を切り出すようにしても良い。

【００５８】ステップ１００２において、単語頻度算定手段908は単語頻度索引903を参照し、総文書数Ｎを求めるとともに検索要求入力手段907で入力された複数の辞書単語Ｗｉ（ｉ＝１、２、・・・、ＮＷ：ＮＷは検索要求に含まれる辞書単語数）に対し、文書Ｄｊ（ｊ＝１、２、・・・、ＮＤ）中での出現頻度ＴＦｉｊと単語Ｗｉの出現する文書数Ｎｉを取得し、頻度スコア算定手段909に出力する。

【００５９】ステップ１００３において、単語出現位置算定手段910は単語出現位置索引904を参照し検索要求入力手段907で入力された複数の単語Ｗｉの文書Ｄｊ中での出現位置を全て求め、単語近接度算定手段911に出力する。

【００６０】ステップ１００４において、頻度スコア算定手段909は単語頻度算定手段908が出力した総文書数ＮＤ、単語Ｗｉの出現する文書数ＮＤｉ、文書Ｄｊにおける単語Ｗｉの出現頻度ＴＦｉｊから、単語頻度による文書Ｄｊと検索要求の合致度合を示す頻度スコアＳＦｊを前記式（１）によって算出する。

【００６１】ステップ１００５において、単語近接度算定手段911は単語出現位置算定手段が出力した文書Ｄｊ中での各単語Ｗｉの出現位置と単語長から、異なる単語の全ての出現位置の組合せＰｋ（ｋ＝１、２、・・・、ＮＰ：ＮＰは異なる単語の全ての出現位置の組合せの数）について２単語の間の距離ＤＳＴｊｋを求め、ステップ１００６において、ＤＳＴｊｋをもとに単語近接度ＮＥｊｋを求める。例えば単語近接度ＮＥｊｋは式（８）を用いて求めることができる。

【００６２】

$$NEjk = 1 / (DSTjk + 1) \quad \cdots (8)$$

【００６３】なお、全ての組合せについて単語近接度を求めるのは計算コストがかかるため、閾値ｄを設け距離ＤＳＴｊｋがｄ以下であるような出現位置の組合せについてのみ計算をしたり、近接度を求める単語ペアをユーザが限定するようにしても良い。

【００６４】ステップ１００７において、近接スコア算定手段912は単語近接度算定手段911の出力した単語近接度ＮＥｊｋにより各文書Ｄｊの近接スコアＳＮｊを式（９）により算出する。

13　　14

$$SNj=\sum_{k}(NEjk) \quad \cdots (9)$$

【００６５】ステップ１００８において、文書スコア算定手段913は頻度スコア算定手段909が出力した頻度スコアＳＦｊと近接スコア算定手段912が出力した近接スコ＊アＳＮｊから文書Ｄｊのスコアｓｊ、すなわち検索要求と文書Ｄｊの合致度合を式（１０）によって算出する。

$$Sj=SFj+SNj\times 定数 \quad \cdots (10)$$

【００６６】このように、近接スコアＳＮｊを用意することにより、検索要求に含まれる異なり単語が互いに接近して出現するような文書のスコアを高くし、優先的に表示させることが可能となる。また、前記式（１０）において定数の値を変化させることにより、単語近接度による優先表示の度合を変えることも可能である。

【００６７】ステップ１００９において、文書順位付け手段914は文書スコア算定手段913が出力した各文書Ｄｊの文書スコアＳｊの値の大きい順に文書を並べ替える。ステップ１０１０において、検索結果表示手段915は文書順位付け手段914の出力から得られたソート済の文書を検索の結果としてユーザに提示する。

【００６８】前記のようにすれば、検索要求に含まれる複数の単語が互いに近くに出現しなければ検索要求として意味をなさないよう場合に、不要な文書が検索結果の上位に来てしまうというような不都合を回避でき、効率的な検索が可能となる。

【００６９】（第５の実施の形態）図９は本発明の第５の実施の形態における文書検索装置の構成を示したブロック図である。図９において、文書検索装置は、検索対象となる文書データ101と、辞書102と、辞書単語の検索対象文書中における出現頻度を格納した単語頻度索引103と、検索対象文書中に現れる単語共起情報を格納した単語共起索引104と、文書データ101から単語頻度情報を抽出し、単語頻度索引103に格納する単語頻度情報抽出手段105と、文書データ101から単語共起情報を抽出し、単語共起索引104に格納する単語共起情報抽出手段106と、ユーザが検索要求を入力するための検索要求入力手段107と、検索要求入力手段107で入力された検索要求に含まれる辞書単語について、単語頻度索引103を調べ当該単語の文書中での出現頻度を求める単語頻度算定手段108と、単語頻度算定手段108で得られた単語頻度をもとに各文書の頻度スコアを求める頻度スコア算定手段109と、検索要求入力手段107で入力された検索要求から単語共起情報を抽出する単語共起情報抽出手段110と、単語共起索引104の内容を参照し、単語共起情報抽出手段110が出力した検索要求に含まれる単語共起関係が、各文書にいくつ現れるかを求める単語共起関係照合手段111と、単語共起関係照合手段111によって得られた検索要求と文書に共通して出現する単語共起関係の度合によって各文書の共起スコアを求める共起スコア算定手段112と、頻度スコア算定手段109の出力と共起スコア算定手段112の出力から文書スコアを算出する文書スコア算定手段113と、スコアの順に文書を並べ替える文書順位付け手段114と、スコアの順に並べられた結果文書を表示する検索結果表示手段115とから構成されている。

【００７０】図１０は本発明の第５の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【００７１】まず、検索の前に、予め単語頻度情報抽出手段105が文書データ101を走査し、総文書数、当該単語の出現文書数とともに単語頻度索引103に出力し、単語共起情報抽出手段106が文書データ101を走査し、各文書内での単語共起情報を求め、単語共起索引104に出力し、索引を作成しておく。単語共起情報としては例えば同一文章内に出現する単語のペアを共起関係にあると判断して抽出する方法や、形態素解析を行なって係受けの関係にある単語のペアを抽出する方法が考えられる。

【００７２】ステップ２０１において、ユーザは検索要求入力手段107によって探したい文書に対する検索要求を文章で入力する。ステップ２０２において、単語頻度算定手段108は単語頻度索引103を参照し、総文書数ＮＤを求めるとともに、検索要求入力手段107で入力された検索要求に含まれる辞書単語Ｗｉ（ｉ＝１、２、・・・、ＮＷ：ＮＷは検索要求に含まれる辞書単語数）に対し、当該単語の文書Ｄｊ（ｊ＝１、２、・・・、ＮＤ）中での出現頻度ＴＦｉｊと当該単語の出現する文書数ＮＤｉを取得し、頻度スコア算定手段109に出力する。

【００７３】ステップ２０３において、頻度スコア算定手段109は単語頻度算定手段108が出力した総文書数ＮＤ、単語Ｗｉの出現する文書数ＮＤｉ、文書Ｄｊにおける単語Ｗｉの出現頻度ＴＦｉｊから、単語頻度による文書Ｄｊと検索要求による合致度合を示す頻度スコアＳＦｊを前記式（１）によって算出する。

【００７４】ステップ２０４において、単語共起情報抽出手段110は検索要求入力手段107で入力された検索要求から、索引作成時と同様の方法によって単語共起関係Ｃｋ（ｋ＝１、２、・・・、ＮＣ：ＮＣは検索要求に含まれる単語共起関係の数）を抽出する。ステップ２０５において、単語共起関係照合手段111は単語共起索引104を参照し、文書Ｄｊに出現する単語共起関係のうち単語共起情報抽出手段110で得られた検索要求に含まれる単語共起関係Ｃｋと一致するものの数ＮＣｊを算出し、共起スコア算定手段112に出力する。

【００７５】ステップ２０６において、共起スコア算定手段112は検索要求と文書の間で一致する単語共起関係の数に基づいて文書Ｄｊの共起スコアＳＣｊを算出する。最も単純な例としては式（１１）のように共起の数をそのまま共起スコアＳＣｊとする。

$$SCj=NCj \quad \cdots (11)$$

【００７６】ステップ２０７において、文書スコア算定手段113は頻度スコア算定手段109の出力した頻度スコアと共起スコア算定手段112の出力した共起スコアから文＊書Ｄｊと検索要求の合致度合を示す文書スコアＳｊを式（１２）によって算出する。

$$Sj=SFj+SCj\times Const \quad \cdots (12)$$

【００７７】ステップ２０８において、文書順位付け手段114は文書スコア算定手段113の出力した各文書Ｄｊの文書スコアＳｊの値の大きい順に文書を並べ替え、ステップ２０９において、検索結果表示手段115が文書順位付け手段114の出力から得られたソート済の文書を検索の結果としてユーザに表示する。

【００７８】前記のようにして、単語頻度だけでなく検索要求と検索対象文書に含まれる単語共起関係を照合し、順位付けに反映させることにより、ユーザの検索意図により近い文書を検索結果の上位に表示することができ、効率的な検索が可能となる。

【００７９】（第６の実施の形態）図１４は本発明の第６の実施の形態における文書検索装置の構成を示したブロック図である。図１４において、文書検索装置は、検索対象となる文書データ1401と、辞書1402と、辞書単語の検索対象文書中における出現頻度を格納した単語頻度索引1403と、検索対象文書中に現れる単語共起情報を格納した単語共起索引1404と、文書データ1401から単語頻度情報を抽出し、単語頻度索引1403に格納する単語頻度情報抽出手段1405と、文書データ1401から単語共起情報を抽出し、単語共起索引1404に格納する単語共起情報抽出手段1406と、ユーザが重要視したいと思う検索要求を入力するための主題検索要求入力手段1407と、ユーザが主題検索要求に比べそれほど重要視しなくても良いと思う検索要求を入力するための副題検索要求入力手段1408と、主題検索要求入力手段1407および副題検索要求入力手段1408で入力された検索要求に含まれる辞書単語について、単語頻度索引1403を調べ当該単語の文書中での出現頻度を求める単語頻度算定手段1409と、単語頻度算定手段1409で得られた単語頻度をもとに各文書の頻度スコアを求める頻度スコア算定手段1410と、主題検索要求入力手段1407および副題検索要求入力手段1408で入力された検索要求から単語共起情報を抽出する単語共起情報抽出手段1411と、単語共起索引1404の内容を参照し、単語共起情報抽出手段1411が出力した検索要求に含まれる単語共起関係が、各文書にいくつ現れるかを求める単語共起関係照合手段1412と、単語共起関係照合手段1412によって得られた検索要求と文書に共通して出現する単語共起関係の数によって各文書の共起スコアを求める共起スコア算定手段1413と、頻度スコア算定手段1410の出力と共起スコア算定手段1413の出力から各文書に対する最終的なスコアを算出する文書スコア算定手段1414と、スコアの順に文書を並べ替える文書順位付け手段1415と、スコアの順に並べられた結果文書を表示する検索結果表示手段1416とから構成される。

【００８０】図１５、図１６、図１７および図１８は本発明の第６の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【００８１】まず、検索の前に予め、単語頻度情報抽出手段1405が文書データ1401を走査し、辞書1402に含まれる辞書単語の各文書内での出現頻度を調べ、総文書数、当該単語の出現文書数とともに単語頻度索引1403に出力し、単語共起情報抽出手段1406が文書データ1401を走査し、各文書内での単語共起情報を求め、単語共起索引1404に出力し、索引を作成しておく。単語共起情報としては例えば同一文章内に出現する単語のペアを共起関係にあると判断して抽出する方法や、形態素解析を行なって係受けの関係にある単語のペアを抽出する方法が考えられる。

【００８２】ステップ１５０１において、ユーザは主題検索要求入力手段1407によって探したい文書に対する検索要求でかつ重視したいものを主題検索要求として入力する。

【００８３】ステップ１５０２において、ユーザは副題検索要求入力手段1408によって主題検索要求に比べそれほど重視しなくてもよいものを副題検索要求として入力する。

【００８４】ステップ１５０３において、単語頻度算定手段1409は単語頻度索引1403を参照し、総文書数ＮＤを求めるとともに、主題検索要求入力手段1407および副題検索要求入力手段1408で入力された検索要求に含まれる辞書単語Ｗｉ（ｉ＝１、２、・・・、ＮＷ：ＮＷは検索要求に含まれる辞書単語数）に対し、当該単語の文書Ｄｊ（ｊ＝１、２、・・・、ＮＤ）中での出現頻度ＴＦｉｊと当該単語の出現する文書数ＮＤｉを取得し、ステップ１５０４において、単語頻度算定手段1409は単語Ｗｉが主題検索要求、副題検索要求のどちらに含まれるかによって重み付けパラメータＷＴｉを選び、頻度スコア算定手段1410に出力する。

【００８５】ステップ１５０５において、頻度スコア算定手段1410は単語頻度算定手段1409が出力した総文書数ＮＤ、単語Ｗｉの出現する文書数ＮＤｉ、文書Ｄｊにおける単語Ｗｉの出現頻度ＴＦｉｊ、単語Ｗｉの重み付けパラメータＷＴｉから、単語頻度による文書Ｄｊと検索要求の合致度合いを示す頻度スコアＳＦｊを前記式（５）によって算出し、文書スコア算定手段1414に出力する。

【００８６】ステップ１５０６において、単語共起情報抽出手段1411は索引作成時と同様の方法によって主題検索要求入力手段1407で入力された主題検索要求から主題

共起関係Ｃｓｋ（ｋ＝１、２、・・・、ＮＣｓ：ＮＣｓは主題検索要求に含まれる単語共起関係の数）を抽出し、単語共起関係照合手段1412に出力する。

【００８７】ステップ１５０７において、単語共起関係照合手段1412は単語共起索引1404を参照し、文書Ｄｊに出現する単語共起関係のうち単語共起情報抽出手段1411で得られた主題共起関係Ｃｓｋと一致するものの数ＮＣｓｊを算出し、共起スコア算定手段1413に出力する。

【００８８】ステップ１５０８において、単語共起情報抽出手段1411は索引作成時と同様の方法によって副題検索要求入力手段1408で入力された副題検索要求から副題共起関係Ｃｆｍ（ｍ＝１、２、・・・、ＮＣｆ：ＮＣｆ＊は副題検索要求に含まれる単語共起関係の数）を抽出し、単語共起関係照合手段1412に出力する。

【００８９】ステップ１５０９において、単語共起関係照合手段1412は単語共起索引1404を参照し、文書Ｄｊに出現する単語共起関係のうち単語共起情報抽出手段1411で得られた副題共起関係Ｃｆｍと一致するものの数ＮＣｆｊを算出し、共起スコア算定手段1413に出力する。

【００９０】ステップ１５１０において、共起スコア算定手段1413は式（１３）に基づいて文書Ｄｊの共起スコアＳＣｊを算出し、文書スコア算定手段1414に出力する。

$$SCj=NCsj\times(NCf+1)+NCfj \quad \cdots(13)$$

【００９１】ステップ１５１１において、文書スコア算定手段1414は式（１４）に基づいて頻度スコアの最大値※と最小値の差ＳＲを算出する。

$$SR=Max(SFj)-Min(SFj) \quad \cdots(14)$$

【００９２】ステップ１５１２において、文書スコア算定手段1414は頻度スコア算定手段1410の出力した頻度スコアと共起スコア算定手段1413の出力した共起スコアから★文書Ｄｊと検索要求との合致度合いを示す文書スコアＳｊを式（１５）によって算出する。

$$Sj=SFj+SCj\times SR \quad \cdots(15)$$

【００９３】ステップ１５１３において、文書順位付け手段1415は文書スコア算定手段1414の出力した各文書Ｄｊの文書スコアＳｊの値の大きい順に文書を並べ替え、ステップ１５１４において検索結果表示手段1416が文書順位付け手段1415の出力から得られたソート済みの文書を検索の結果としてユーザに表示する。前記のようにして、主題検索要求と副題検索要求という重要視する度合の異る検索要求を受け付け、検索要求と文書の合致度合いを判定する基準として、主題共起関係＞副題共起関係＞主題単語頻度＞副題単語頻度、の順に優先することにより、ユーザの検索意図により近い文書を検索結果の上位に表示することができ、高精度で効率的な検索が可能となる。

【００９４】（第７の実施の形態）図１９は本発明の第７の実施の形態における文書検索装置の構成を示したブロック図である。図１９において、文書検索装置は、検索対象となる文書データ1901と、辞書1902と、辞書単語の検索対象文書中における出現頻度をフィールド毎に格納したフィールド別単語頻度索引1903と、検索対象文書中に現れる単語共起情報をフィールド毎に格納したフィールド別単語共起索引1904と、文書データ1901から単語頻度情報を抽出し、フィールド別単語頻度索引1903に格納する単語頻度情報抽出手段1905と、文書データ1901から単語共起情報を抽出し、フィールド別単語共起索引1904に格納する単語共起情報抽出手段1906と、ユーザが検索要求を入力するための検索要求入力手段1907と、検索要求入力手段1907で入力された検索要求に含まれる辞書単語について、フィールド別単語頻度索引1903を調べ当該単語の文書中でのフィールド毎の出現頻度を求めるフィールド別単語頻度算定手段1908と、フィールド別単語頻度算定手段1908で得られた単語頻度をもとに各文書のフィールド毎の頻度スコアを求めるフィールド別頻度スコア算定手段1909と、検索要求入力手段1907で入力された検索要求から単語共起情報を抽出する単語共起情報抽出手段1910と、フィールド別単語共起索引1904の内容を参照し、単語共起情報抽出手段1910が出力した検索要求に含まれる単語共起関係が、各文書の各フィールドにいくつ現れるかを求めるフィールド別単語共起関係照合手段1911と、フィールド別単語共起関係照合手段1911によって得られた検索要求と文書の各フィールドに共通して出現する単語共起関係の数によって各文書のフィールド毎の共起スコアを求めるフィールド別共起スコア算定手段1912と、各フィールドのスコアをどの程度文書の順位付けにスコアに反映させるかという割合をユーザが入力するためのフィールド割合入力手段1913と、フィールド別頻度スコア算定手段1909の出力とフィールド別共起スコア算定手段1912の出力とフィールド割合入力手段1913の出力から各文書に対する最終的なスコアを算出する文書スコア算定手段1914と、スコアの順に文書を並べ替える文書順位付け手段1915と、スコアの順に並べられた結果文書を表示する検索結果表示手段1916とから構成される。

【００９５】図２０、図２１、図２２および図２３は本発明の第７の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【００９６】まず、検索の前に予め、単語頻度情報抽出手段1905が文書データ1901を走査し、辞書1902に含まれる辞書単語の各文書内でのフィールド毎の出現頻度を調べ、総文書数、当該単語の出現文書数とともにフィールド別単語頻度索引1903出力し、単語共起情報抽出手段19

06が文書データ1901を走査し、各文書内でのフィールド毎の単語共起情報を求め、フィールド別単語共起索引1904に出力し、索引を作成しておく。単語共起情報としては例えば同一文章内に出現する単語のペアを共起関係にあると判断して抽出する方法や、形態素解析を行なって係受けの関係にある単語のペアを抽出する方法が考えられる。

【００９７】ステップ２００１において、ユーザは検索要求入力手段1907によって探したい文書に対する検索要求を入力する。

【００９８】ステップ２００２において、ユーザはフィールド割合入力手段1913によってフィールドＦｍ（ｍ＝１、２、・・・、ＮＦ：ＮＦは総フィールド数）のスコアを順位付けに反映させる割合Ｒｍを入力する。

【００９９】ステップ２００３において、フィールド別単語頻度算定手段1908はフィールド別単語頻度索引1903を参照し、総文書数ＮＤを求めるとともに、検索要求入力手段1907で入力された検索要求に含まれる辞書単語Ｗｉ（ｉ＝１、２、・・・、ＮＷ：ＮＷは検索要求に含まれる辞書単語数）に対し、当該単語の文書Ｄｊ（ｊ＝１、２、・・・、ＮＤ）のフィールドＦｍ中での出現頻度ＴＦｉｊｍとフィールドＦｍに当該単語の出現する文書数ＮＤｉｍを取得し、フィールド別頻度スコア算定手段1909に出力する。

＊

$$SCjm=NCjm \quad \cdots (16)$$

※コアの最大値と最小値の差ＳＲを算出する。

【０１０４】ステップ２００８において、文書スコア算定手段1914は式（１７）に基づいてフィールド別頻度ス※

$$SR=Max(SFjm)-Min(SFjm) \quad \cdots (17)$$

【０１０５】ステップ２００９において、文書スコア算定手段1914はフィールド別頻度スコア算定手段1909の出力したフィールド毎の頻度スコアＳＦｊｍとフィールド別共起スコア算定手段1912の出力したフィールド毎の共★

★起スコアＳＣｊｍとフィールド割合入力手段で入力されたスコア配分割合Ｒｍから文書Ｄｊと検索要求との合致度合いを示す文書スコアＳｊを式（１８）によって算出する。

$$Sj=\sum_{m}((SFjm+SCjm\times SR)\times Rm) \quad \cdots (18)$$

【０１０６】ステップ２０１０において文書順位付け手段1915は文書スコア算定手段1914の出力した各文書Ｄｊの文書スコアＳｊの値の大きい順に文書を並べ替え、ステップ２０１１において検索結果表示手段1916が文書順位付け手段1915の出力から得られたソート済みの文書を検索の結果としてユーザに表示する。

【０１０７】前記のようにして、ユーザが検索対象フィールドのスコア配分の割合を変化させられるようにすることにより、ユーザの検索意図を柔軟に表現することができ、効率的な検索が可能となる。

【０１０８】（第８の実施の形態）図２４は本発明の第８の実施の形態における文書検索装置の構成を示したブロック図である。図２４において、文書検索装置は、検索対象となる文書データ2401と、辞書2402と、辞書単語の検索対象文書中における出現頻度を格納した単語頻度索引2403と、検索対象文書中に現れる単語共起情報を格納した単語共起索引2404と、文書データ2401から単語頻度情報を抽出し、単語頻度索引2403に格納する単語頻度情報抽出手段2405と、文書データ2401から単語共起情報を抽出し、単語共起索引2404に格納する単語共起情報抽出手段2406と、ユーザが検索要求を入力するための検索要求入力手段2407と、検索要求入力手段2407で入力された検索要求に含まれる辞書単語について、単語頻度索引2403を調べ当該単語の文書中での出現頻度を求める単語頻度算定手段2408と、単語頻度算定手段2408で得られた単語頻度をもとに各文書の頻度スコアを求める頻度スコア算定手段2409と、単語頻度索引2403を調べ、検索要求入力手段2407で入力された検索要求に含まれる辞書単語が、各文書中にいくつ出現するのかを求める出現語数算定手段2410と、出現語数算定手段2411で得られた出現語数をもとに各文書の出現語数スコアを求める出現語数スコア算定手段2411と、検索要求入力手段2407で入力され

た検索要求から単語共起情報を抽出する単語共起情報抽出手段2412と、単語共起索引2404の内容を参照し、単語共起情報抽出手段2412が出力した検索要求に含まれる単語共起関係が、各文書にいくつ現れるかを求める単語共起関係照合手段2413と、単語共起関係照合手段2413によって得られた検索要求と文書に共通して出現する単語共起関係の数によって各文書の共起スコアを求める共起スコア算定手段2414と、頻度スコア算定手段2409の出力と出現語数スコア算定手段2411の出力と共起スコア算定手段2414の出力から各文書に対する最終的なスコアを算出する文書スコア算定手段2415と、スコアの順に文書を並べ替える文書順位付け手段2416と、スコアの順に並べられた結果文書を表示する検索結果表示手段2417とから構成される。

【０１０９】図２５、図２６、図２７、図２８および図２９は本発明の第８の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【０１１０】まず、検索の前に予め、単語頻度情報抽出手段2405が文書データ2401を走査し、辞書2402に含まれる辞書単語の各文書内での出現頻度を調べ、総文書数、当該単語の出現文書数とともに単語頻度索引2403に出力し、単語共起情報抽出手段2406が文書データ2401を走査し、各文書内での単語共起情報を求め、単語共起索引2404に出力し、索引を作成しておく。単語共起情報としては例えば同一文章内に出現する単語のペアを共起関係にあると判断して抽出する方法や、形態素解析を行なって係受けの関係にある単語のペアを抽出する方法が考えられる。

【０１１１】ステップ２５０１において、ユーザは検索要求入力手段2407によって探したい文書に対する検索要求を入力する。

【０１１２】ステップ２５０２において、単語頻度算定手段2408は単語頻度索引2403を参照し、総文書数ＮＤを求めるとともに、検索要求入力手段2407で入力された検索要求に含まれる辞書単語Ｗｉ（ｉ＝１、２、・・・、ＮＷ：ＮＷは検索要求に含まれる辞書単語数）に対し、当該単語の文書Ｄｊ（ｊ＝１、２、・・・、ＮＤ）中での出現頻度ＴＦｉｊと当該単語の出現する文書数ＮＤｉを取得し、頻度スコア算定手段2409に出力する。

【０１１３】ステップ２５０３において、頻度スコア算定手段2409は単語頻度算定手段2408が出力した総文書数＊ＮＤ、単語Ｗｉの出現する文書数ＮＤｉ、文書Ｄｊにおける単語Ｗｉの出現頻度ＴＦｉｊから、単語頻度による文書Ｄｊと検索要求の合致度合いを示す頻度スコアＳＦｊを前記式（１）によって算出し、文書スコア算定手段2415に出力する。

【０１１４】ステップ２５０４において、出現語数算定手段2410は単語頻度索引2403を参照し、検索要求入力手段2407で入力された検索要求に含まれる辞書単語Ｗｉのうち、文書Ｄｊに出現する単語の数ＮＡｊを算出し、出現語数スコア算定手段2411に出力する。

【０１１５】ステップ２５０５において、出現語数スコア算定手段2411は出現語数算定手段2410の出力した出現語数ＮＡｊにもとづいて、文書Ｄｊの出現語数スコアを前記式（６）によって算出し、出現語数スコア算定手段2411に出力する。

【０１１６】ステップ２５０６において、単語共起情報抽出手段2412は索引作成時と同様の方法によって検索要求入力手段2407で入力された検索要求から共起関係Ｃｋ（ｋ＝１、２、・・・、ＮＣ：ＮＣは検索要求に含まれる単語共起関係の数）を抽出し、単語共起関係照合手段2413に出力する。

【０１１７】ステップ２５０７において、単語共起関係照合手段2413は単語共起索引2404を参照し、単語共起情報抽出手段2404で得られた各単語共起関係Ｃｋが出現する文書を求め、単語共起関係Ｃｋのうちで文書Ｄｊに出現するものの数ＮＣｊを算出し、共起スコア算定手段2414に出力する。

【０１１８】ステップ２５０８において、共起スコア算定手段2414は前記式（１１）に基づいて文書Ｄｊの共起スコアＳＣｊを算出し、文書スコア算定手段2415に出力する。

【０１１９】ステップ２５０９において、文書スコア算定手段2415は前記式（１４）に基づいて頻度スコアの最大値と最小値の差ＳＲを算出する。

【０１２０】ステップ２５１０において、文書スコア算定手段2415は頻度スコア算定手段2409の出力した頻度スコアＳＦｊと出現語数スコア算定手段2411の出力した出現語数スコアＳＡｊと共起スコア算定手段2414の出力した共起スコアＳＣｊから文書Ｄｊと検索要求との合致度合いを示す文書スコアＳｊを式（１９）によって算出する。

$$Sj = SFj + (SAj + SCj \times NW) \times SR \quad \cdots (19)$$

【０１２１】ステップ２５１１において、文書順位付け手段2416は文書スコア算定手段2415の出力した各文書Ｄｊの文書スコアＳｊの値の大きい順に文書を並べ替え、ステップ２５１２において検索結果表示手段2417が文書順位付け手段2416の出力から得られたソート済みの文書を検索の結果としてユーザに表示する。

【０１２２】前記のようにして、検索要求と文書の合致度合いを判定する基準として、単語頻度だけでなく、共起関係、出現語数を採り入れ、共起関係＞出現語数＞単語頻度、の順に優先することにより、ユーザの検索意図により近い文書を検索結果の上位に表示することができ、高精度で効率的な検索が可能となる。

【０１２３】

【発明の効果】以上のように本発明の文書検索装置においては、優先度の異なる複数の検索要求を入力するための複数の検索要求入力手段を設けることにより、また、

検索対象文書のフィールド毎の索引情報を持ち、順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段を設けることにより、また、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語数算定手段を設けることにより、また、検索対象文書中の単語出現位置を索引に持ち、検索要求に含まれる複数の単語の検索対象文書中での出現位置が隣接もしくは比較的近いかどうかを調べる単語近接度算定手段を設けることにより、また、検索対象文書中の単語共起情報を索引に持ち、複数の検索要求入力手段および検索要求に含まれる単語共起関係が検索対象文書に現れるか否かを調べる単語共起関係照合手段を設けることにより、また、検索対象文書中の単語出現頻度と単語共起情報をフィールド毎に索引に持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べるフィールド別単語共起関係照合手段を設けることにより、また、検索対象文書中の単語出現頻度と単語共起情報を索引に持ち、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語算定手段および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べる単語共起関係照合手段を設けることにより、ユーザの検索要求を柔軟に受け付け、検索、順位付けを行なうことにより、ゴミの少ない高精度な検索結果が得られ、結果の絞り込みを含めた総合的な検索時間が短縮可能な文書検索装置が得られるものである。

【図面の簡単な説明】

【図１】本発明の第１の実施の形態における文書検索装置の構成を示すブロック図、

【図２】本発明の第１の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図３】本発明の第２の実施の形態における文書検索装置の構成を示すブロック図、

【図４】本発明の第２の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図５】本発明の第３の実施の形態における文書検索装置の構成を示すブロック図、

【図６】本発明の第３の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図７】本発明の第４の実施の形態における文書検索装置の構成を示すブロック図、

【図８】本発明の第４の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図９】本発明の第５の実施の形態における文書検索装置の構成を示すブロック図、

【図１０】本発明の第５の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図１１】従来の文書検索装置の構成を示すブロック図、

【図１２】従来の文書検索装置の検索の手順を示す流れ図、

【図１３】従来の文書検索装置の検索の例を示す図、

【図１４】本発明の第６の実施の形態における文書検索装置の構成を示すブロック図、

【図１５】本発明の第６の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図１６】本発明の第６の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図１７】本発明の第６の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図１８】本発明の第６の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図１９】本発明の第７の実施の形態における文書検索装置の構成を示すブロック図、

【図２０】本発明の第７の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図２１】本発明の第７の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図２２】本発明の第７の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図２３】本発明の第７の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図２４】本発明の第８の実施の形態における文書検索装置の構成を示すブロック図、

【図２５】本発明の第８の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図２６】本発明の第８の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図２７】本発明の第８の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図２８】本発明の第８の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図２９】本発明の第８の実施の形態における文書検索装置の検索の手順を示す流れ図である。

【符号の説明】

101、301、501、701、901、1101、1401、1901、2401　文書データ

102、302、502、702、902、1102、1402、1902、2402　辞書

103、303、503、703、903、1103、1403、2403　単語頻度索引

104、1404、2404　単語共起索引

105、304、504、705、905、1104、1405　単語頻度情報抽出手段

1905、2405　単語頻度情報抽出手段

106、1406、1906、2406　単語共起情報抽出手段

107、505、707、907、1105、1907、2407　検索要求入力手段

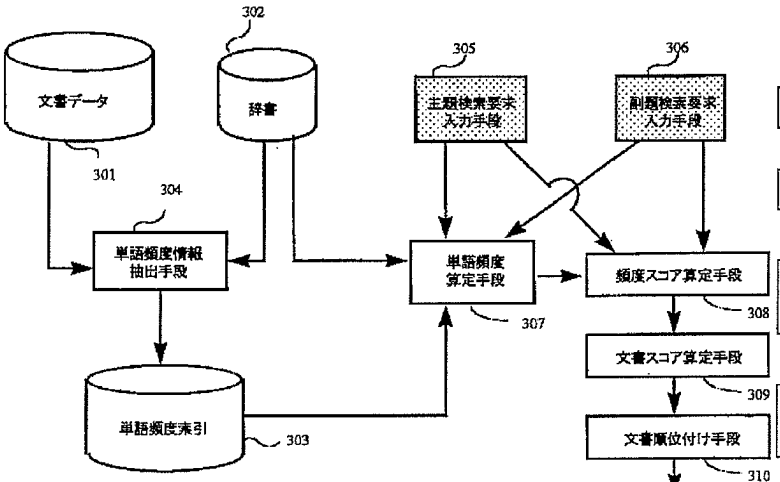
25

108、307、708、908、1106、1409、2408　単語頻度算定手段
109、308、709、909、1107、1410、2409　頻度スコア算定手段
110、1406、1411、1906、1910、2406、2412　単語共起情報抽出手段
111、1412、2413　単語共起関係照合手段
112、1413、2414　共起スコア算定手段
113、309、509、712、913、1108、1414　文書スコア算定手段
1914、2415　文書スコア算定手段
114、310、510、713、914、1109、1415　文書順位付け手段
1915、2416　文書順位付け手段
115、311、511、714、915、1110、1416　検索結果表示手段

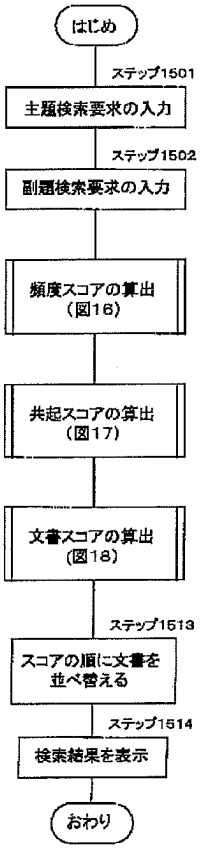
26

＊1916、2417　検索結果表示手段
305、1407　主題検索要求入力手段
306、1408　副題検索要求入力手段
506、1913　フィールド割合入力手段
507、1908　フィールド別単語頻度算定手段
508、1909　フィールド別頻度スコア算定手段
710、2410　出現語数算定手段
711、2411　出現語数スコア算定手段
904　単語出現位置索引
906　単語出現位置情報抽出手段
910　単語出現位置算定手段
911　単語近接度算定手段
912　近接スコア算定手段
1911　フィールド別単語共起関係照合手段
1912　フィールド別共起スコア算定手段

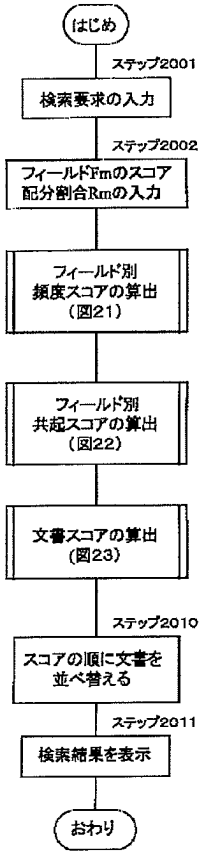
【図１】



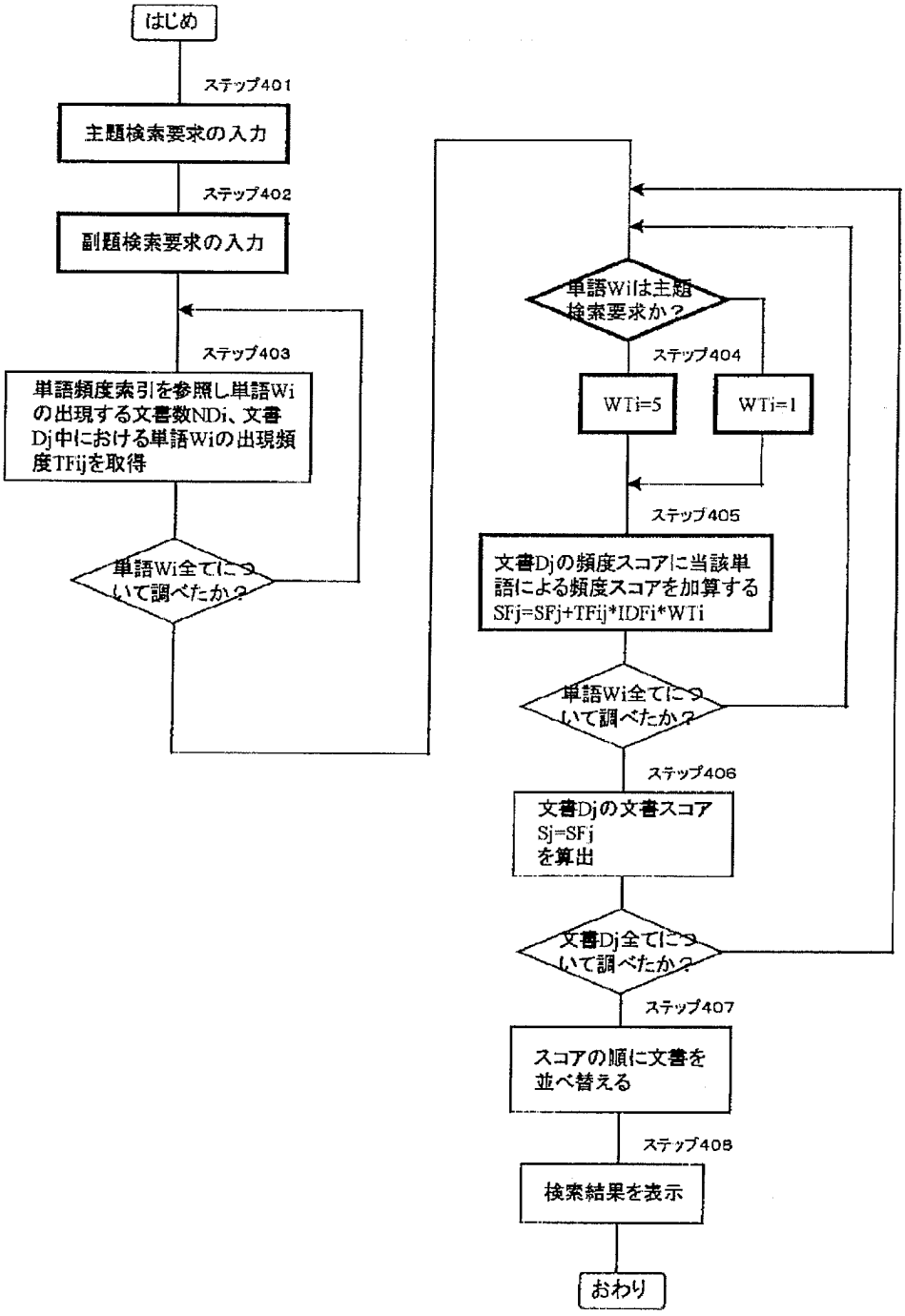
【図１５】



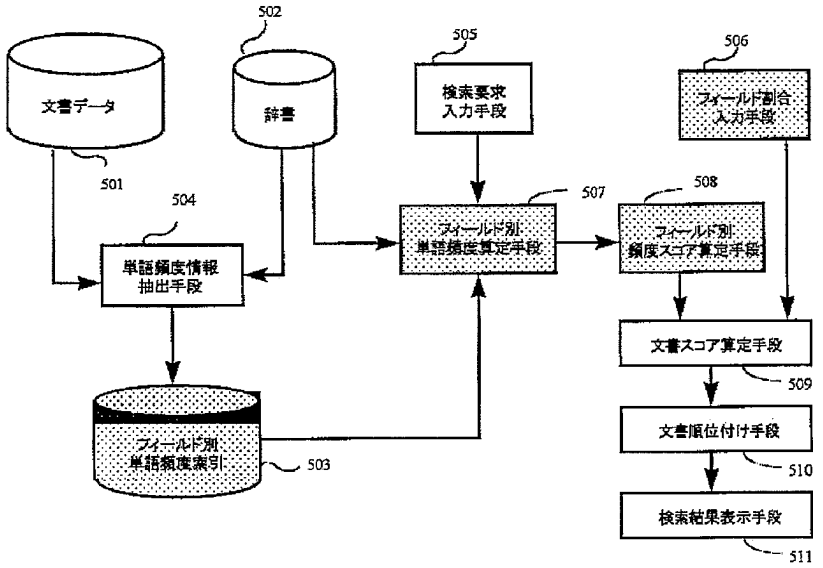
【図２０】



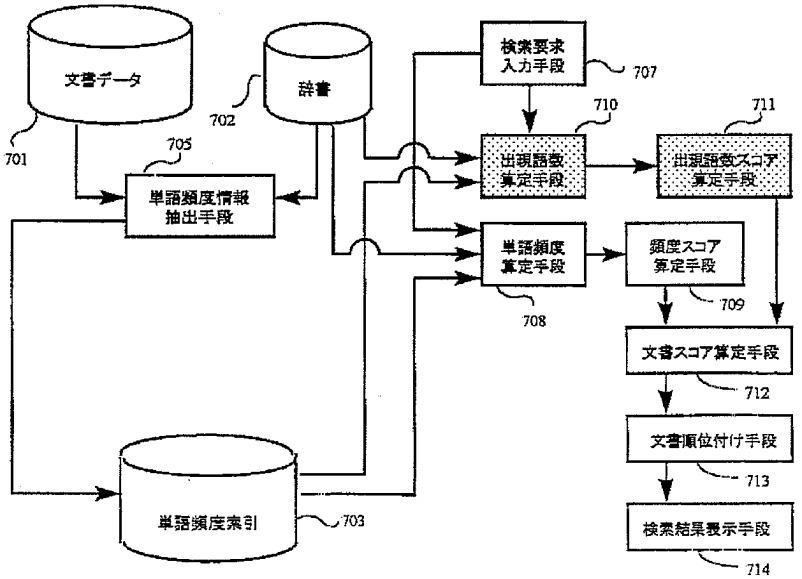
【図２】


はじめ
ステップ401
主題検索要求の入力
ステップ402
副題検索要求の入力
ステップ403
単語頻度索引を参照し単語Wiの出現する文書数NDi、文書Dj中における単語Wiの出現頻度TFijを取得
単語Wi全てについて調べたか？
単語Wiは主題検索要求か？
ステップ404
WTi=5
WTi=1
ステップ405
文書Djの頻度スコアに当該単語による頻度スコアを加算する
SFj=SFj+TFij*IDFi*WTi
単語Wi全てについて調べたか？
ステップ406
文書Djの文書スコア
Sj=SFj
を算出
文書Dj全てについて調べたか？
ステップ407
スコアの順に文書を並べ替える
ステップ408
検索結果を表示
おわり

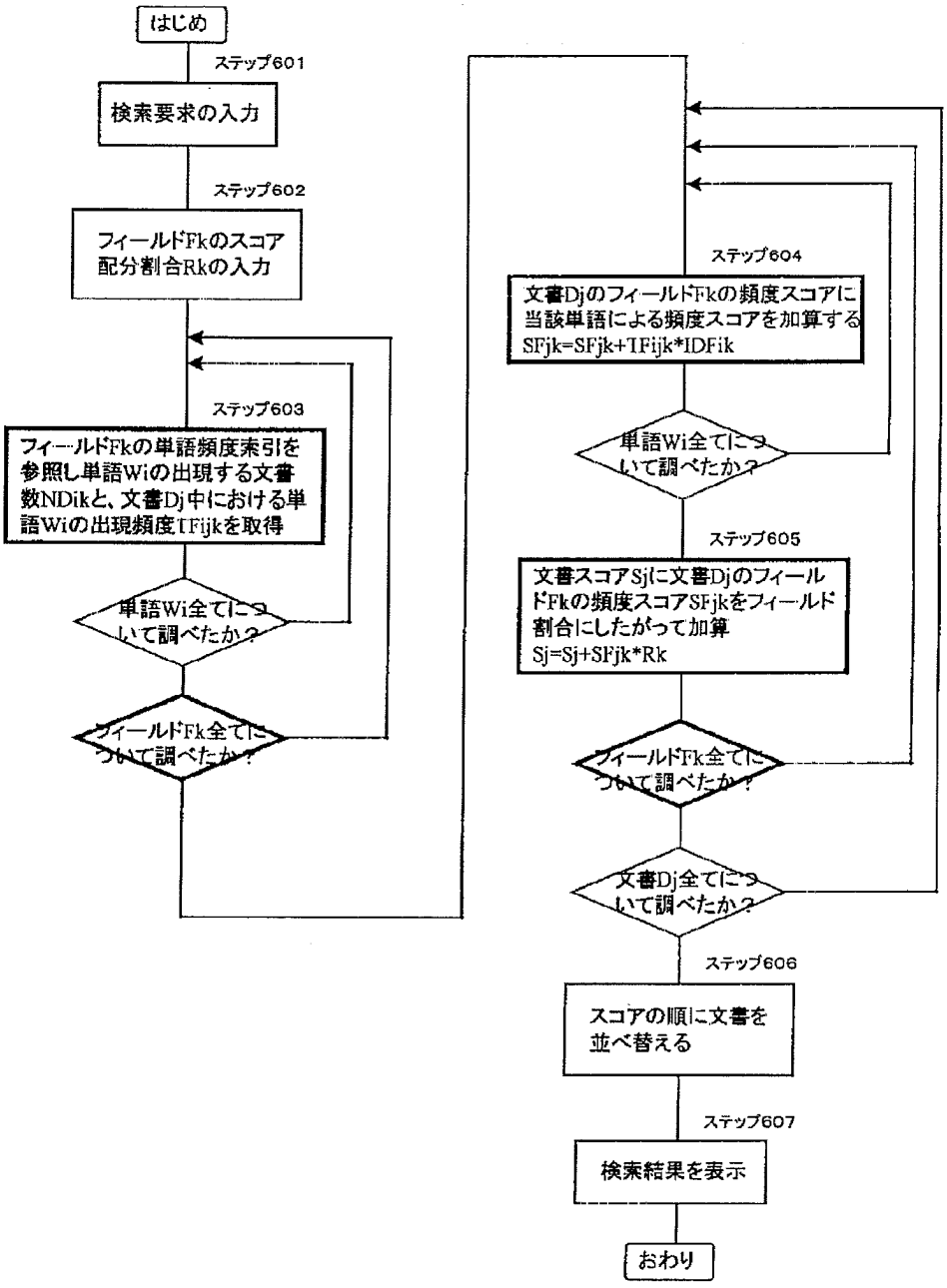
【図３】


文書データ
501
辞書
502
検索要求
入力手段
505
フィールド割合
入力手段
506
単語頻度情報
抽出手段
504
フィールド別
単語頻度索引
503
フィールド別
単語頻度算定手段
507
フィールド別
頻度スコア算定手段
508
文書スコア算定手段
509
文書順位付け手段
510
検索結果表示手段
511

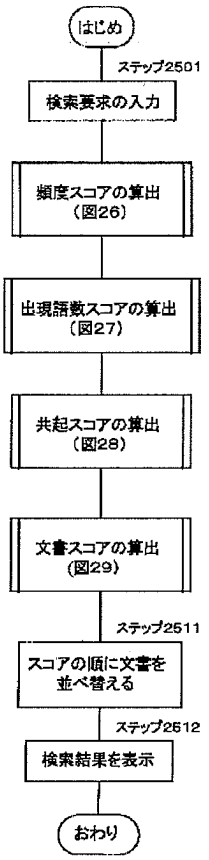
【図５】


文書データ
701
辞書
702
検索要求
入力手段
707
出現語数
算定手段
710
出現語数スコア
算定手段
711
単語頻度情報
抽出手段
705
単語頻度
算定手段
708
頻度スコア
算定手段
709
文書スコア算定手段
712
文書順位付け手段
713
検索結果表示手段
714
単語頻度索引
703

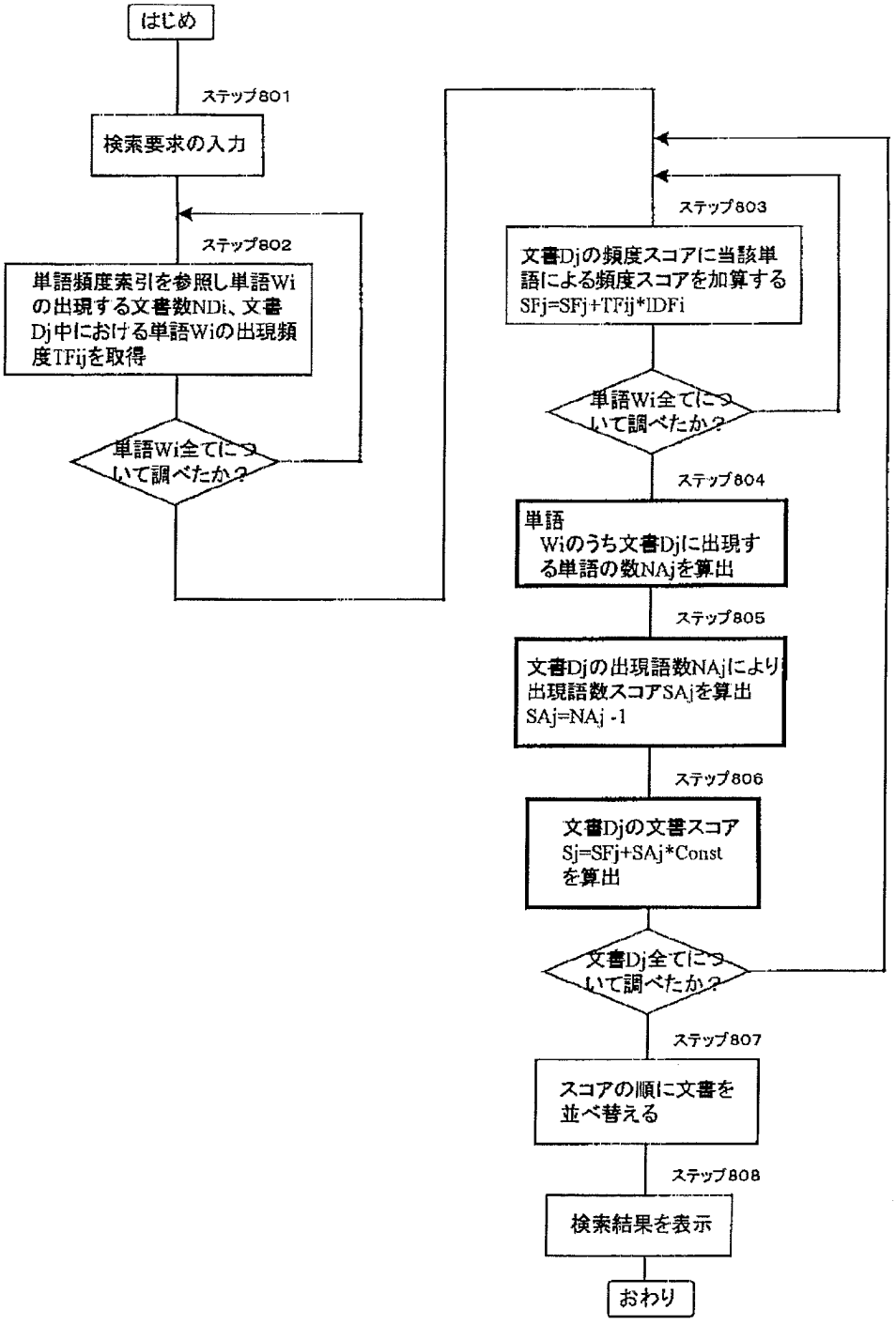
【図４】


はじめ
ステップ601
検索要求の入力
ステップ602
フィールドFkのスコア配分割合Rkの入力
ステップ603
フィールドFkの単語頻度索引を参照し単語Wiの出現する文書数NDikと、文書Dj中における単語Wiの出現頻度TFijkを取得
単語Wi全てについて調べたか？
フィールドFk全てについて調べたか？
ステップ604
文書Djのフィールドの頻度スコアに当該単語による頻度スコアを加算する
SFjk=SFjk+TFijk*IDFik
単語Wi全てについて調べたか？
ステップ605
文書スコアSjに文書Djのフィールドの頻度スコアSFjkをフィールド割合にしたがって加算
Sj=Sj+SFjk*Rk
フィールドFk全てについて調べたか？
文書Dj全てについて調べたか？
ステップ606
スコアの順に文書を並べ替える
ステップ607
検索結果を表示
おわり

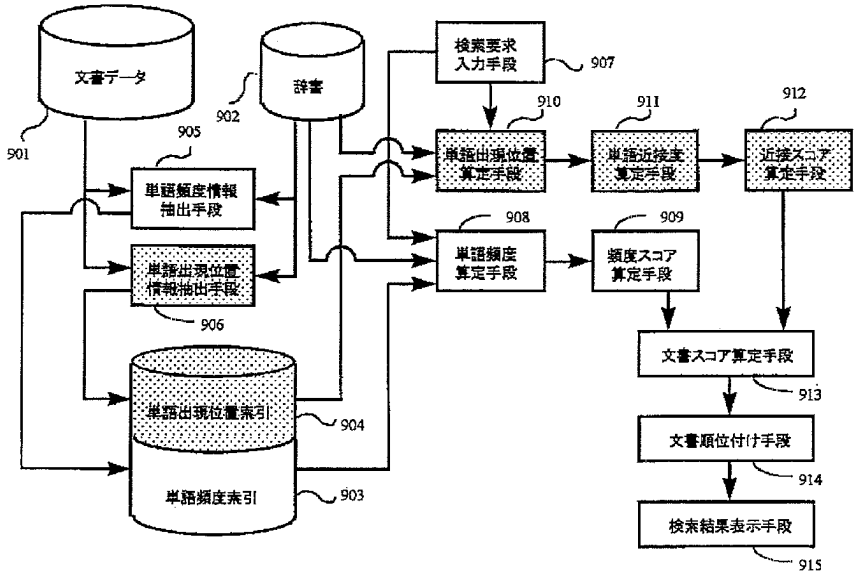
【図２５】


はじめ
ステップ2501
検索要求の入力
頻度スコアの算出（図26）
出現語数スコアの算出（図27）
共起スコアの算出（図28）
文書スコアの算出（図29）
ステップ2511
スコアの順に文書を並べ替える
ステップ2512
検索結果を表示
おわり

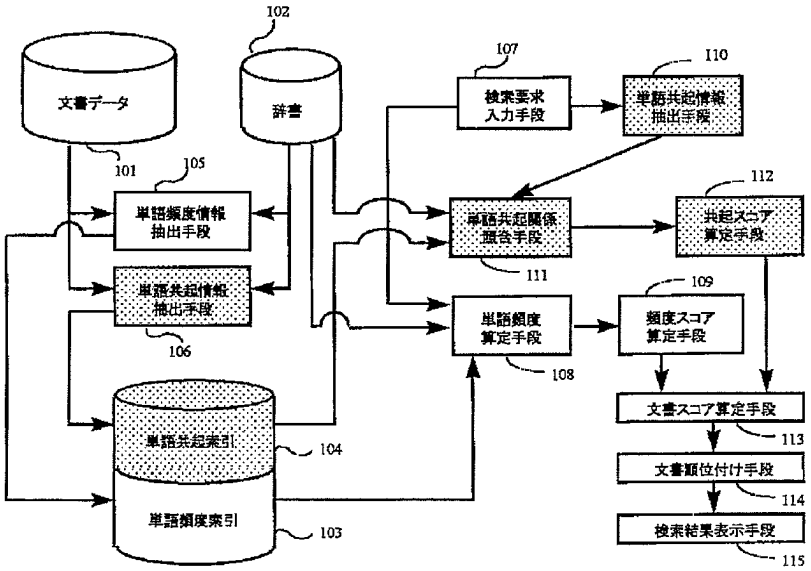
【図６】


はじめ
ステップ801
検索要求の入力
ステップ802
単語頻度索引を参照し単語Wiの出現する文書数NDi、文書Dj中における単語Wiの出現頻度TFijを取得
単語Wi全てについて調べたか？
ステップ803
文書Djの頻度スコアに当該単語による頻度スコアを加算する
SFj=SFj+TFij*IDFi
単語Wi全てについて調べたか？
ステップ804
単語Wiのうち文書Djに出現する単語の数NAjを算出
ステップ805
文書Djの出現語数NAjにより出現語数スコアSAjを算出
SAj=NAj -1
ステップ806
文書Djの文書スコアSj=SFj+SAj*Const を算出
文書Dj全てについて調べたか？
ステップ807
スコアの順に文書を並べ替える
ステップ808
検索結果を表示
おわり

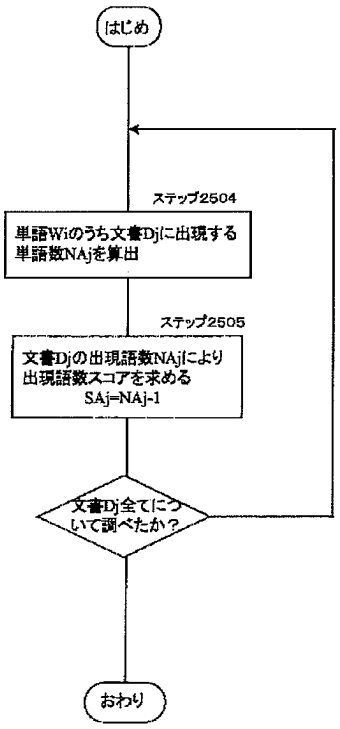
【図７】


文書データ
901
辞書
902
検索要求
入力手段
907
910
911
912
単語出現位置
算定手段
単語近接度
算定手段
近接スコア
算定手段
905
単語頻度情報
抽出手段
単語出現位置
情報抽出手段
906
908
909
単語頻度
算定手段
頻度スコア
算定手段
単語出現位置索引
904
単語頻度索引
903
文書スコア算定手段
913
文書順位付け手段
914
検索結果表示手段
915

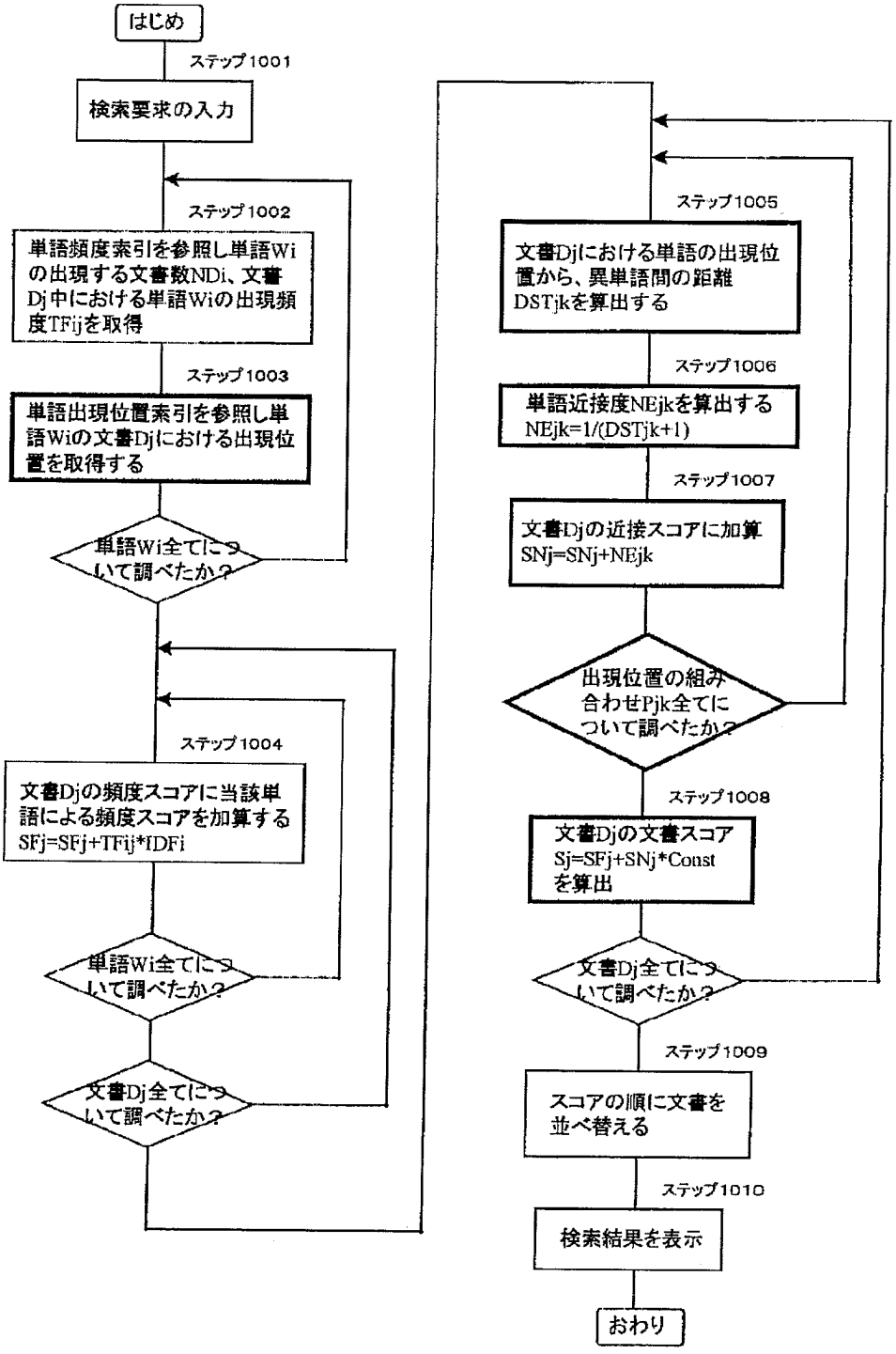
【図９】


102
文書データ
101
辞書
107
検索要求
入力手段
110
単語共起情報
抽出手段
105
単語頻度情報
抽出手段
単語共起関係
算定手段
112
共起スコア
算定手段
単語共起情報
抽出手段
106
111
109
単語頻度
算定手段
頻度スコア
算定手段
108
単語共起索引
104
単語頻度索引
103
文書スコア算定手段
113
文書順位付け手段
114
検索結果表示手段
115

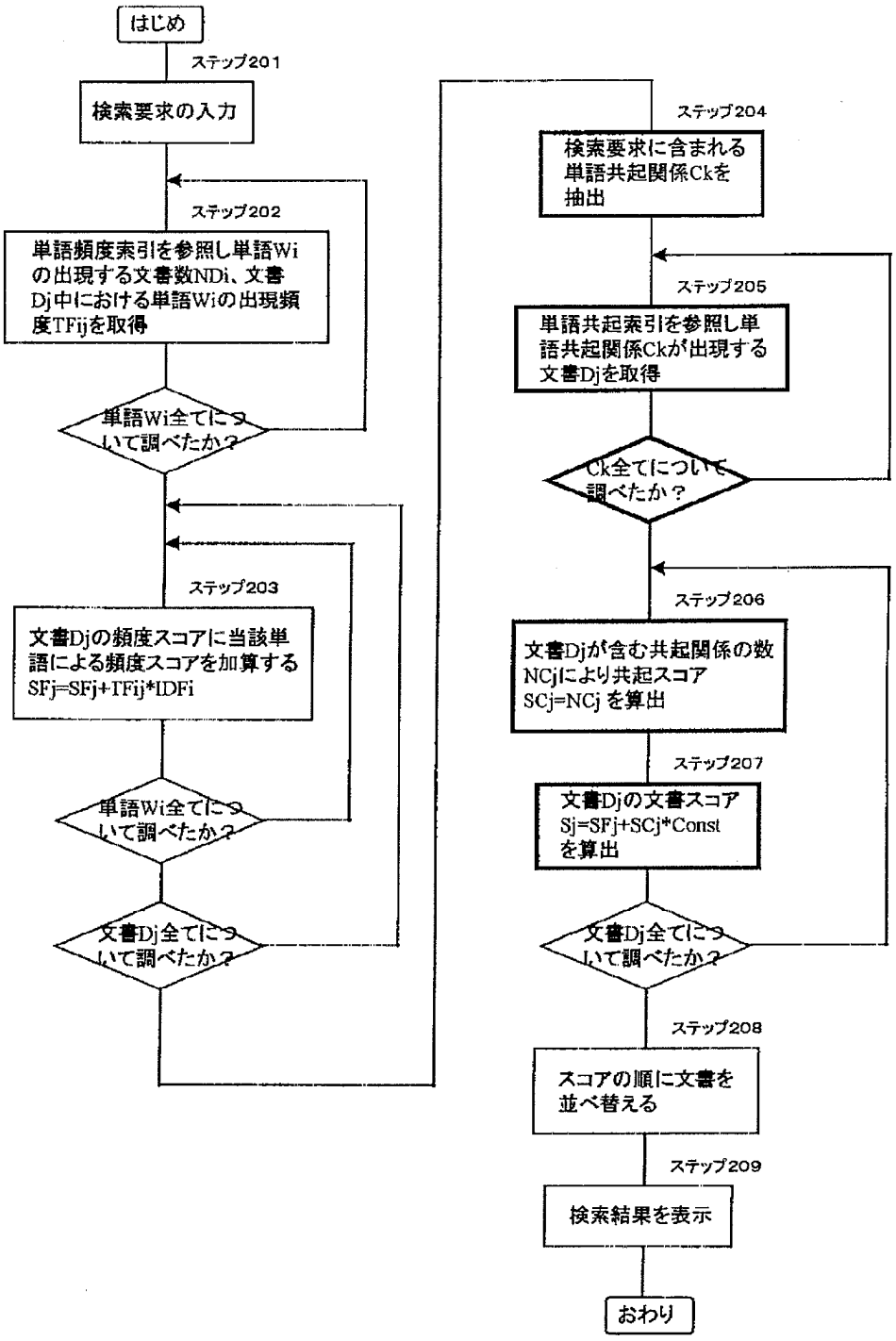
【図２７】


はじめ
ステップ2504
単語Wiのうち文書Djに出現する
単語数NAjを算出
ステップ2505
文書Djの出現語数NAjにより
出現語数スコアを求める
SAj=NAj-1
文書Dj全てについて調べたか？
おわり

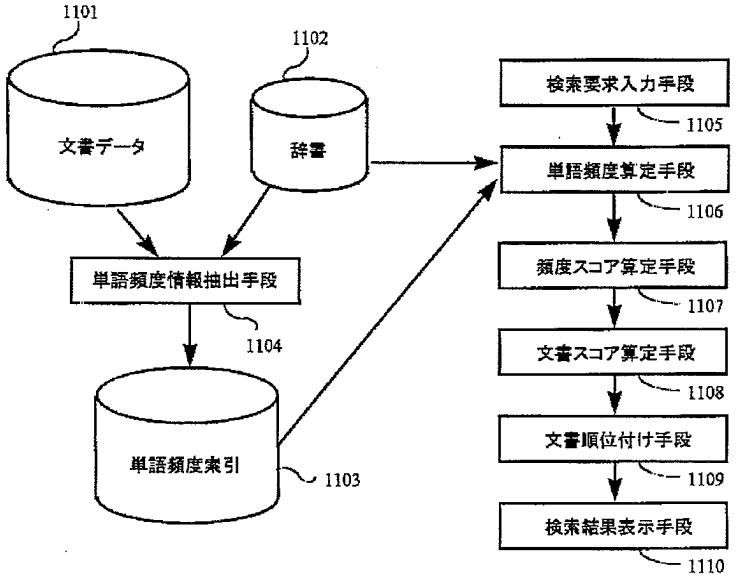
【図８】


はじめ
ステップ1001
検索要求の入力
ステップ1002
単語頻度索引を参照し単語Wiの出現する文書数NDi、文書Dj中における単語Wiの出現頻度TFijを取得
ステップ1003
単語出現位置索引を参照し単語Wiの文書Djにおける出現位置を取得する
単語Wi全てについて調べたか？
ステップ1004
文書Djの頻度スコアに当該単語による頻度スコアを加算する
SFj=SFj+TFij*IDFi
単語Wi全てについて調べたか？
文書Dj全てについて調べたか？
ステップ1005
文書Djにおける単語の出現位置から、異単語間の距離DSTjkを算出する
ステップ1006
単語近接度NEjkを算出する
NEjk=1/(DSTjk+1)
ステップ1007
文書Djの近接スコアに加算
SNj=SNj+NEjk
出現位置の組み合わせPjk全てについて調べたか？
ステップ1008
文書Djの文書スコア
Sj=SFj+SNj*Const
を算出
文書Dj全てについて調べたか？
ステップ1009
スコアの順に文書を並べ替える
ステップ1010
検索結果を表示
おわり

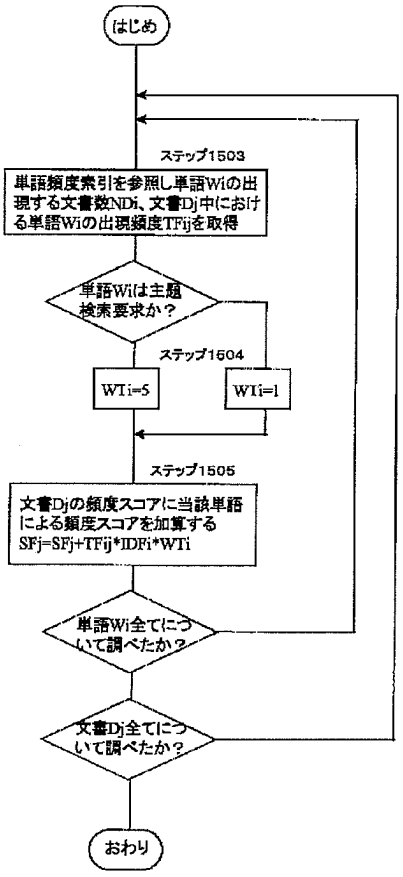
【図１０】


はじめ
ステップ201
検索要求の入力
ステップ202
単語頻度索引を参照し単語Wiの出現する文書数NDi、文書Dj中における単語Wiの出現頻度TFijを取得
単語Wi全てについて調べたか？
ステップ203
文書Djの頻度スコアに当該単語による頻度スコアを加算する SFj=SFj+TFij*IDFi
単語Wi全てについて調べたか？
文書Dj全てについて調べたか？
ステップ204
検索要求に含まれる単語共起関係Ckを抽出
ステップ205
単語共起索引を参照し単語共起関係Ckが出現する文書Djを取得
Ck全てについて調べたか？
ステップ206
文書Djが含む共起関係の数NCjにより共起スコア SCj=NCj を算出
ステップ207
文書Djの文書スコア Sj=SFj+SCj*Const を算出
文書Dj全てについて調べたか？
ステップ208
スコアの順に文書を並べ替える
ステップ209
検索結果を表示
おわり

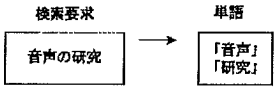
【図１１】

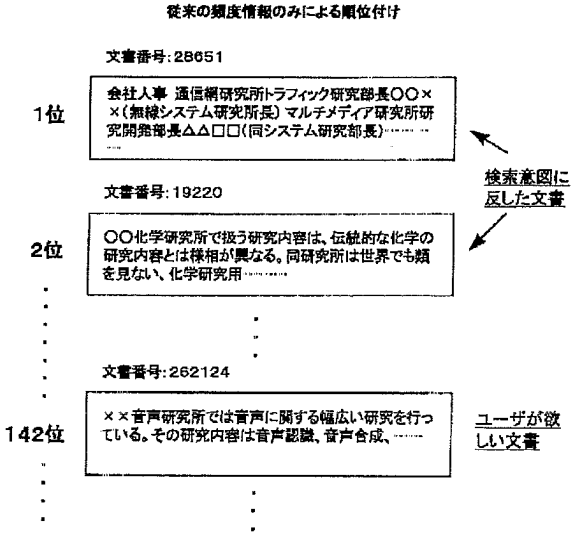

1101
1102
文書データ
辞書
単語頻度情報抽出手段
1104
単語頻度索引
1103
検索要求入力手段
1105
単語頻度算定手段
1106
頻度スコア算定手段
1107
文書スコア算定手段
1108
文書順位付け手段
1109
検索結果表示手段
1110

【図１６】

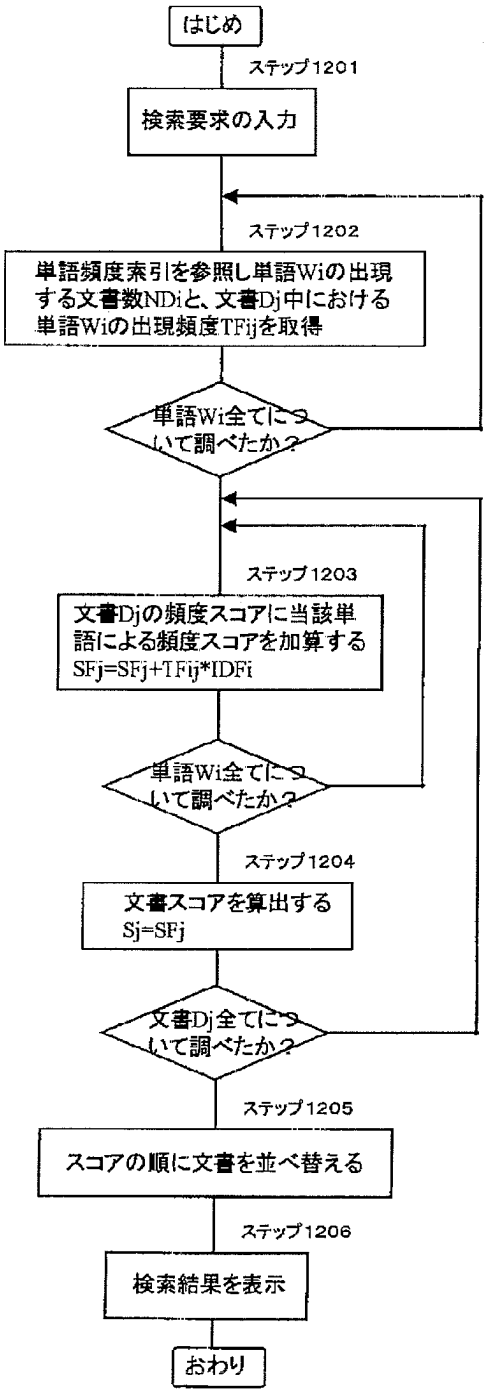

はじめ
ステップ1503
単語頻度索引を参照し単語Wiの出現する文書数NDi、文書Dj中における単語Wiの出現頻度TFijを取得
単語Wiは主題検索要求か？
ステップ1504
WIi=5
WIi=1
ステップ1505
文書Djの頻度スコアに当該単語による頻度スコアを加算するSFj=SFj+TFij*IDFi*WTi
単語Wi全てについて調べたか？
文書Dj全てについて調べたか？
おわり

【図１３】

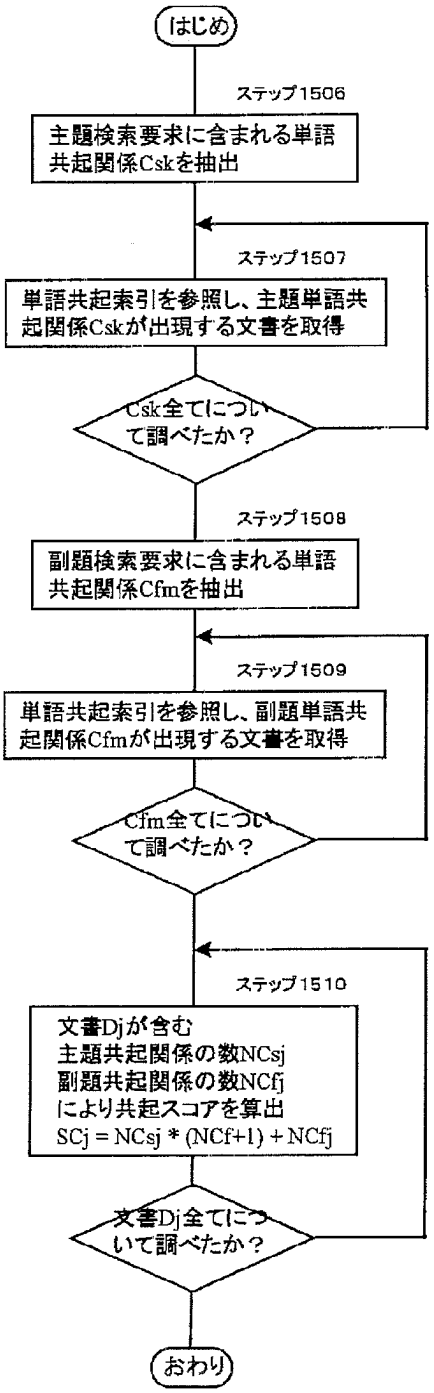

検索要求
単語
音声の研究
「音声」
「研究」


従来の頻度情報のみによる順位付け
文書番号:28651
1位
会社人事 通信網研究所トラフィック研究部長○○×
×(無線システム研究所長) マルチメディア研究所研究開発部長△△□□(同システム研究部長)……
検索意図に反した文書
文書番号:19220
2位
○○化学研究所で扱う研究内容は、伝統的な化学の研究内容とは様相が異なる。同研究所は世界でも類を見ない、化学研究用……
文書番号:262124
142位
××音声研究所では音声に関する幅広い研究を行っている。その研究内容は音声認識、音声合成、……
ユーザが欲しい文書

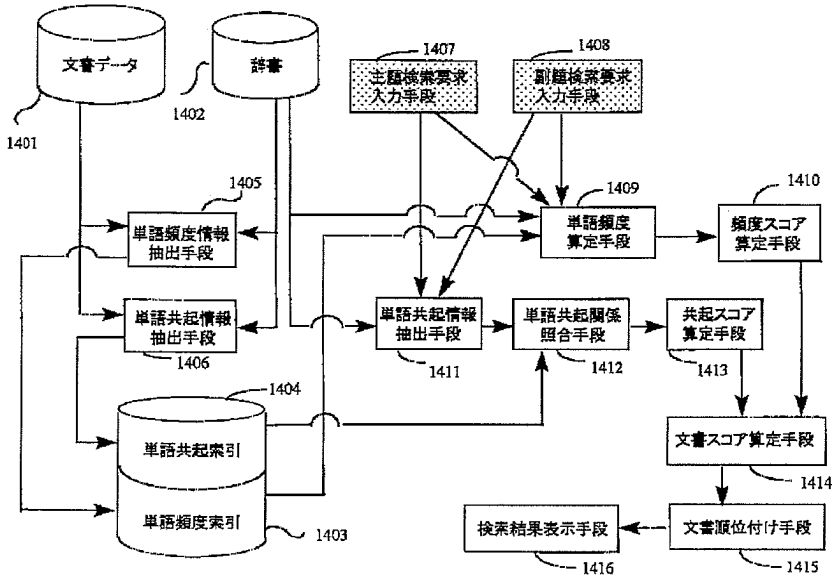
【図１２】


はじめ
ステップ1201
検索要求の入力
ステップ1202
単語頻度索引を参照し単語Wiの出現する文書数NDiと、文書Dj中における単語Wiの出現頻度TFijを取得
単語Wi全てについて調べたか？
ステップ1203
文書Djの頻度スコアに当該単語による頻度スコアを加算する
SFj=SFj+TFij*IDFi
単語Wi全てについて調べたか？
ステップ1204
文書スコアを算出する
Sj=SFj
文書Dj全てについて調べたか？
ステップ1205
スコアの順に文書を並べ替える
ステップ1206
検索結果を表示
おわり

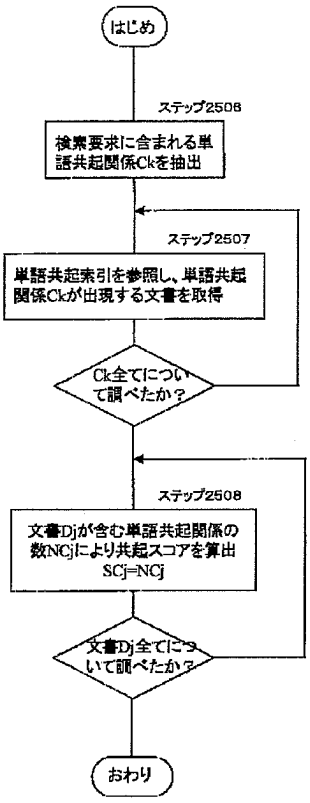
【図１７】


はじめ
ステップ1506
主題検索要求に含まれる単語共起関係Cskを抽出
ステップ1507
単語共起索引を参照し、主題単語共起関係Cskが出現する文書を取得
Csk全てについて調べたか？
ステップ1508
副題検索要求に含まれる単語共起関係Cfmを抽出
ステップ1509
単語共起索引を参照し、副題単語共起関係Cfmが出現する文書を取得
Cfm全てについて調べたか？
ステップ1510
文書Djが含む
主題共起関係の数NCsj
副題共起関係の数NCfj
により共起スコアを算出
SCj = NCsj * (NCf+1) + NCfj
文書Dj全てについて調べたか？
おわり

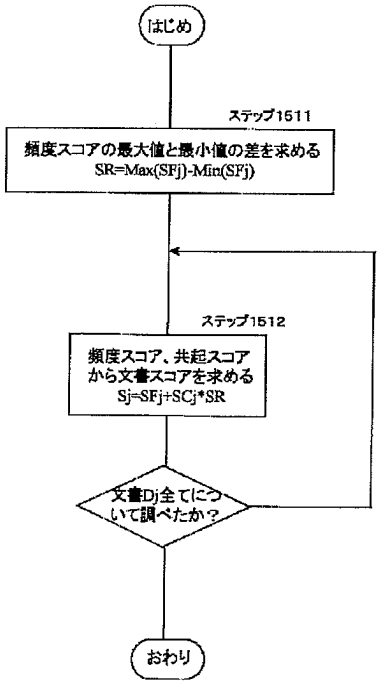
【図１４】



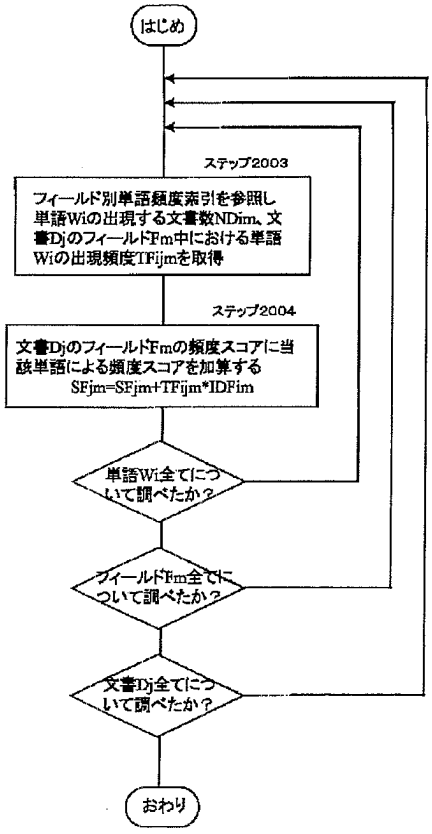
【図２８】



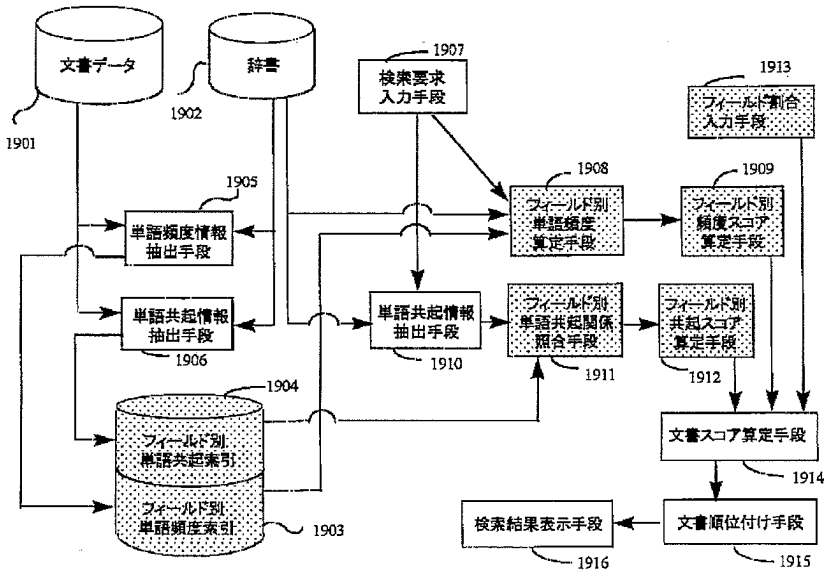
【図１８】



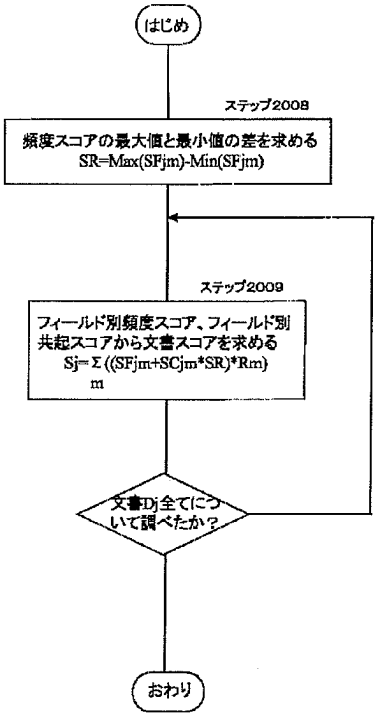
【図２１】



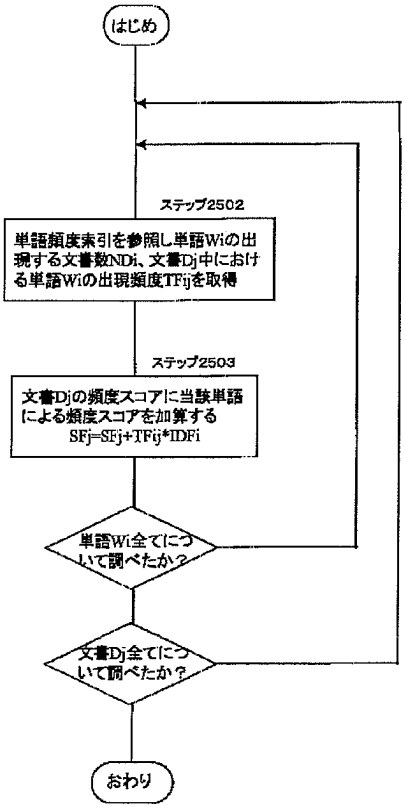
【図１９】



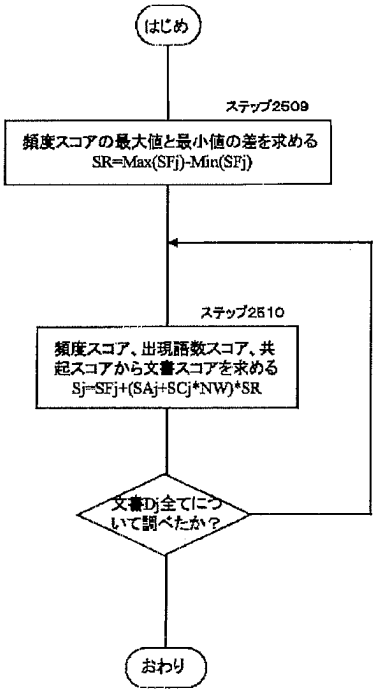
【図２３】



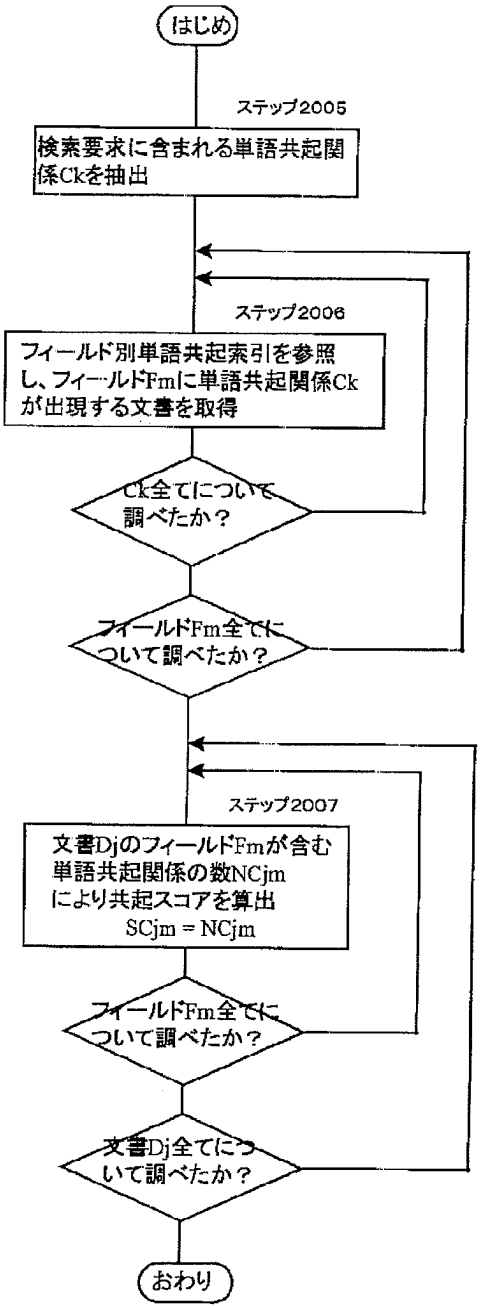
【図２６】



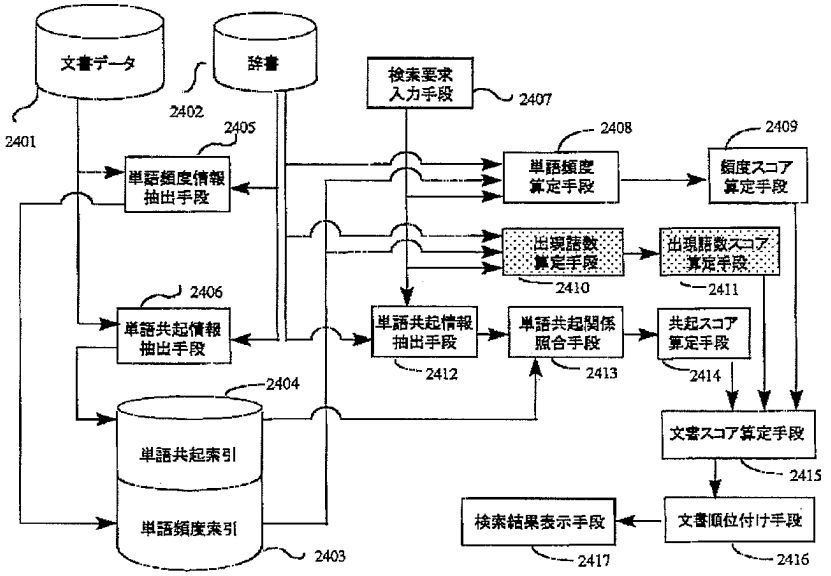
【図２９】



【図２２】


はじめ
ステップ2005
検索要求に含まれる単語共起関係Ckを抽出
ステップ2006
フィールド別単語共起索引を参照し、フィールドFmに単語共起関係Ckが出現する文書を取得
Ck全てについて調べたか？
フィールドFm全てについて調べたか？
ステップ2007
文書Djのフィールドfmが含む単語共起関係の数NCjmにより共起スコアを算出
SCjm = NCjm
フィールドFm全てについて調べたか？
文書Dj全てについて調べたか？
おわり

【図２４】



フロントページの続き

(72)発明者　佐藤　光弘
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　野本　昌子
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　安川　秀樹
大阪府門真市大字門真1006番地　松下電器産業株式会社内